Research Evaluation, 30(1), 2021, 112–126
doi: 10.1093/reseval/rvaa032
Advance Access Publication Date: 25 December 2020
Article

The role of metrics in peer assessments

Liv Langfeldt *, Ingvild Reymert and Dag W. Aksnes

Nordic Institute for Studies in Innovation, Research and Education (NIFU), P.O. Box 2815 Tøyen, N- 0608 Oslo, Norway

*Corresponding author. Email: liv.langfeldt@nifu.no.

Abstract

Metrics on scientific publications and their citations are easily accessible and are often referred to in assessments of research and researchers. This paper addresses whether metrics are considered a legitimate and integral part of such assessments. Based on an extensive questionnaire survey in three countries, the opinions of researchers are analysed. We provide comparisons across academic fields (cardiology, economics, and physics) and contexts for assessing research (identifying the best research in their field, assessing grant proposals and assessing candidates for positions). A minority of the researchers responding to the survey reported that metrics were reasons for considering something to be the best research. Still, a large majority in all the studied fields indicated that metrics were important or partly important in their review of grant proposals and assessments of candidates for academic positions. In these contexts, the citation impact of the publications and, particularly, the number of publications were emphasized. These findings hold across all fields analysed, still the economists relied more on productivity measures than the cardiologists and the physicists. Moreover, reviewers with high scores on bibliometric indicators seemed more frequently (than other reviewers) to adhere to metrics in their assessments. Hence, when planning and using peer review, one should be aware that reviewers—in particular reviewers who score high on metrics—find metrics to be a good proxy for the future success of projects and candidates, and rely on metrics in their evaluation procedures despite the concerns in scientific communities on the use and misuse of publication metrics.

Key words: peer review; research quality; bibliometric indicators; metrics; research fields

1. Introduction

Research organizations, funding agencies, national authorities and other organizations rely on peer assessments in their research evaluations. Peer assessments, in turn, may (partly) rely on metrics on scientific publications and their citations. In recent decades, such bibliometric indicators have become more easily accessible and have been used more in the evaluation of research. This raises the question of how such metrics impact what is perceived as good research, i.e. the notions of research quality. This paper addresses whether metrics are considered a legitimate and integral part of the assessment of research, explore the role of metrics in different review contexts and fields of research, and discuss implications for research evaluation and policy.

The use of metrics has a long history, dating back more than 100 years (De Bellis 2009). With the creation of the Science Citation Index by Eugene Garfield in 1961, new possibilities for quantitative studies of scientific publishing emerged, including analyses of how often the articles had been referred to or cited in subsequent scientific literature. Initially, the potential of bibliometrics within science policy was only seen by a few individuals (Martin 1996). Later, research evaluation became an important area of application of bibliometric analyses. Today, indicators or metrics are applied for a variety of purposes and have permeated many aspects of the research system (Abbott et al. 2010; Aksnes, Langfeldt and Wouters 2019). For example, metrics have long been provided to peer reviewers in research evaluations, such as in national research assessments and institutional reviews (Lewison, Cottrell and Dixon 1999; Wilsdon et al. 2015). Nowadays, individual applicants may be requested to provide standardized Curriculum Vitae (CVs) that include citations rates when applying for grants[1], and metrics may also play an important role in hiring and promotion processes (Stephan, Veugelers and Wang 2017).

The use of bibliometric indicators has been more common in the natural and medical sciences than in the social sciences and humanities (Moed 2005). This may be due to the fact that the latter areas

are less covered by standard bibliometric databases like Web of Science or Scopus (Aksnes and Sivertsen 2019). They also have different communication practices with more publications in books and in national languages, and a slow accumulation of citations (although there is large heterogeneity at the level of disciplines). However, studies have shown that—even in the social sciences—it has become a common practice for researchers to include metrics in their CVs, applications for promotions, and grant applications (Haddow and Hammarfelt 2019).

There are different types of metrics and a large variety of indicators (for an overview, see e.g. Ball 2017). By metrics, in this paper we refer to publication-based indictors wherein three types are investigated: *productivity/number of publications*, *scientific impact/citations* and the *impact factor of journals* where the publications appear. The most basic is the number of publications, which typically is regarded as an indirect measure of the volume of knowledge production. Citations and citation indicators, on the other hand, are commonly applied as proxies for the impact (or influence) of the research, one of the constituents of the concept of scientific quality (Aksnes, Langfeldt and Wouters 2019). One of the most popular and well-known bibliometric indicators is the journal impact factor (JIF), which is a measure of the frequency the average article in a journal has been cited. The impact factor is often regarded as an indicator of the significance and prestige of a journal (Glänzel and Moed 2002). To what extent bibliometric measures can be used as proxies for these dimensions of research activities is, however, a matter of debate. Particularly, this issue has been addressed with respect to citation indicators, and many studies have, over the years, been carried out in order to assess their validity and appropriateness as performance measures (Aksnes, Langfeldt and Wouters 2019).

The use of metrics has always been controversial and is a key debate in research evaluation practices (Wilsdon et al. 2015). There are many examples of their misuse, and potentially negative impacts of metrics upon the research system have received increased attention (Weingart 2005; de Rijcke et al. 2016). General concerns about metrics being used when assessing individual researchers are expressed in key documents, such as the Leiden Manifesto (Hicks et al. 2015), which contains 10 principles for the appropriate measurement of research performance, as well as the San Francisco Declaration on Research Assessment[2] (DORA), which intends to prevent the practice of using the journal impact factor as a surrogate measure of the quality of individual articles.

2. Backdrop and research questions

Despite the large amount of attention devoted to these issues, there are few empirical studies investigating researchers' use of metrics in different evaluation processes and to which extent their own position, age, gender, and bibliometric performance affect this use. For example, publication metrics are not part of the criteria appearing in a recent review of studies of the criteria used to assess grant applications (Hug and Aeschbach 2020). The present study analyses the use of metrics when assessing the past achievements of applicants for positions and grants. Based on a questionnaire survey, different types of metrics are addressed: journal impact factors, citation indicators, and indicators on number of publications. To enable the exploration of possible diversity in the use of metrics, this study covers three main fields: cardiology, economics and physics, in three countries (Netherlands, Norway, and Sweden). These fields are different in terms of how knowledge production is organized and valued (Whitley 1984), and in the way they relate to metrics. Moreover, there are notable differences between these countries when it comes to the role of metrics in national research policy. As an introduction, we therefore give some brief background information on these issues.

Economics is a field wherein journal rankings have long traditions and are highly influential. Such rankings play a role, for example, in evaluating the performance of economics departments and in hiring processes (Kalaitzidakis, Mamuneas and Stengos 2011; Gibson, Anderson and Tressler 2014). Many rankings exist (Bornmann, Butz and Wohlrabe 2018). In particular, much importance is attached to publishing in the so-called 'Top Five' journals of economics (Hylmö 2018), and a study by Heckman and Moktan (2018) showed that publishing in these journals greatly increases the probability of the author(s) receiving tenure and promotion.

In medicine, the journal impact factor has, over a long time, been a very popular indicator and has been used for purposes such as those described above, as well as for ranking lists delineating where scientists ought to submit their publications. There are many reports on this issue, covering medicine more generally (Brown 2007; Sousa and Hendriks 2007; Allen 2010; Hammarfelt and Rushforth 2017) and cardiology more specifically (van der Wall 2012; Coats and Shewan 2015; Loomba and Anderson 2018). According to van der Wall (2012), publishing in journals with an impact factor below five is considered a signal of 'mediocre scientific quality' in some institutions and departments.

In physics, on the other hand, the use of impact factors appears to be less prevalent compared with medicine, although there is a journal hierarchy whereby certain journals, such as *Physical Review Letters*, are considered to be among the most prestigious (Bollen, Rodriguez and Van De Sompel 2006). Moreover, there are some very large journals, such as the *Physical Review* series, and several physics journals are among the world's largest journals in terms of publication counts.

The three academic fields also have different publication profiles, which may be expected to influence the respondents' views on metrics. The average number of publications per researcher is generally higher in medicine and the natural sciences when compared to the humanities and the social sciences. A study by Piro, Aksnes and Rorstad (2013) found that, in economics, researchers (on average) published 4.4 publications during a four-year period, compared with 5.3 for clinical medicine and 9.5 for physics. However, the average for physics is highly influenced by individuals having extremely high publication output due to their participation in articles with hyper-authorship (articles with several hundred authors, Cronin 2001). Such papers appear in high energy physics, particularly when related to the European Organization for Nuclear Research (CERN). According to Birnholtz (2008), hyper-authorship makes it difficult to identify the roles of individual contributors, which may undermine authorship as the traditional currency of science with respect to performance assessments and career advancement.

This study includes data from multiple countries and also, at the national level, there are differences which might influence the respondents' views on metrics. In Norway, there is a performance-based funding model whereby bibliometric indicators are applied for the allocation of funding across institutions. The system allocating funding to Norwegian universities is based on (among other things) publication indicators where publication channels are divided into quality levels (Sivertsen 2017). In Sweden, governmental

institutional funding has previously been granted partly based on bibliometric indicators on publications and citations in Web of Science (Hammarfelt 2018). While these systems are designed to work on an overall national level, they are sometimes applied at lower levels as well, such as faculties, departments, and individual researchers. This is documented in an evaluation of the Norwegian model (Aagaard 2015). In Sweden, several universities have applied the Norwegian publication indicator to allocate resources within institutions (Hammarfelt 2018). In the Netherlands, institutional funding is not linked to bibliometric measurement systems (Wilsdon et al. 2015; Jonkers and Zacharewicz 2016), but there are still research assessments (organized every sixth year). Here, evaluations are made by expert panels, which may use qualitative as well as quantitative indicators to assess research groups or programmes (Wilsdon et al. 2015).[3] In such evaluations, panels consisting of a few members are often asked to assess the research of several hundred individuals, wherein the total research output may encompass more than a thousand publications.

As for the use of bibliometrics for the kind of assessments addressed in this article (assessments of research funding applications and hiring processes), there is no systematic overview of practices across organizations or countries. Moreover, reviewers are based across organizations and countries, and their propensity to use metrics in assessments may or may not be shaped by the use of metrics in national systems for performance-based funding and research assessments. In sum, how this may vary across countries is not obvious.

More generally, there are at least three separate reasons why peer reviewers may opt to use metrics as (part of) their basis for assessments of grant proposals or of candidates for academic positions. Evaluation processes involve categorization—that is, examining the characteristics of the entities to be assessed and locating them in one or more hierarchies (Lamont 2012), and metrics may thus be helpful in several ways. First, metrics are easily accessible, and they ease the review task in terms of the time and effort required (Espeland and Sauder 2007: 17). Rather than spending time reading the applicants' publications, a reviewer may get an impression by looking up bibliometric indicators (citations counts, h-index, journal impact factor or similar). Second, metrics may be used because the reviewers find them to be good—or fair—proxies for research quality or research performance. They may, for example, find that in their field the best research is published in the highest-ranking journals (as these tend to have the strictest review processes), or that highly cited papers are those that prove most important for the development of the field (by introducing new and valuable knowledge), whereas non-cited papers seldom prove to have any significance. They may also find that comparing applicants based on such indicators provides a more objective, fair and reliable basis for assessments compared to peer assessments that are not informed by such indicators.[4] Finally, the use of metrics may be explicitly encouraged by those organizing the review. It may be part of review criteria and guidelines, and the organizer may provide the metrics to be used.[5]

Similar types of reasons for introducing metrics (availability; good/fair proxies; encouraged from outside) may motivate research and funding organizations. At the organizational level, metrics provide easily accessible information about applicants, they may be perceived as highly relevant and impartial, having the potential to reduce biases in peer assessments, and may also be encouraged by national authorities. Moreover, successful sister organizations using metrics may serve as role models.[6]

Concerning the reasons for reviewers' individual use of metrics, the first and the last types of reasons are obviously present both in grant reviews and reviews of candidates for academic positions: metrics are easily available and at least some funding agencies and research organizations encourage their use. The second type of reason, that metrics are perceived as being good proxies for research quality or research performance, is more uncertain and may vary substantially by field. Moreover, as peer reviewers have discretionary power and the basis of their judgements is not monitored, it may be a necessary condition that the reviewers perceive metrics to be an adequate basis for assessments. If they find metrics to be good proxies, they can be expected to use them, regardless of whether they are encouraged in the guidelines and/or provided to them. Conversely, if they perceive metrics to be an inadequate tool for evaluation, they may disregard guidelines encouraging their use and/or the metrics provided to them.

Against this background, this study addresses two main research questions:

a. To what extent are metrics part of researchers' notion of good research?
b. To what extent are metrics used in reviews of research proposals and in reviews of candidates for academic positions?

The first question was investigated by asking the respondents to characterize the best research in their field, and whether high journal impact factors and many citations are among these characteristics. To answer the second question, we studied the respondents' emphases for assessments of research proposals and candidates for academic positions. This issue was investigated for two types of indicators: publication productivity and citation impact. We aim to understand why some researchers are more apt to rely on metrics in their assessments, and explore how the use of metrics varies between field of research and other background characteristics.

Based on previous studies, we expect views on metrics to be diverse, both within fields and within countries (Aksnes and Rip 2009; Wilsdon et al. 2015; Söderlind and Geschwind 2020)[7]. In a survey to researchers who reviewed grant proposals for the Research Council of Norway (RCN) (including reviewers in all fields of research, most of them from European countries apart from Norway), some commented that they would like the RCN to provide standardized metrics to the reviewers, while others stated that the RCN should try to minimize the weight put on metrics (Langfeldt and Scordato 2016).

3. Data and methods

This paper draws on data from a web survey which explored varying notions and conditions of good research. The survey was filled out by researchers in physics, economics and cardiology in the Netherlands, Norway and Sweden. The three fields belong to different parts of science (the social sciences, the natural sciences and the medical sciences), and as noted above, they differ in publication profiles and in the use of metrics.

3.1 Sampling and response rates

The invited survey sample included all researchers active in the aforementioned three fields at the most relevant universities in the

three countries, as defined by Web of Science data and journal classification. For this, a three step sampling strategy was used: in step one, we used journal categories to identify institutions with a minimum number of articles in the relevant journal categories in the period 2011–2016 (Web of Science (WoS) categories: 'Physics', 'Astronomy & Astrophysics'; 'Economics'; 'Cardiac & cardiovascular systems'). In step two, the websites of these institutions were searched for relevant organizational units to include in the survey, and we generated lists of personnel in relevant academic positions (including staff members, post-docs and researchers—not including PhD students, adjunct positions, guest researchers or administrative and technical personnel). Some departments also had research groups in other disciplines than the one selected. In these cases, we removed the personnel found in the non-relevant groups. In step three, we added people (at the selected institutions) prevailing with a minimum number of WoS publications in the field, regardless of which department/unit they were affiliated with. For economics, a limit of at least five WoS publications (in 2011–2016) was used. In the case of cardiology and physics, where the publication frequency (and co-authorship) is higher, a minimum of 10 publications was used.[8] In this way, we combined two sampling strategies in order to obtain a comprehensive sample: Based on the organizations' websites, we identified the full scope of researchers within a department/division (step two), and based on WoS categories, we identified those who publish in the field (step three).

The web survey yielded viable samples of researchers for each of the three fields; in total, there were 1621 replies[9] (32.7% response of those invited to the survey). The response rates varied substantially by country: 49.1% in Norway, 38.7% in Sweden, and 19.9% in the Netherlands. Response rates also varied somewhat by field (25.8% in cardiology, 31.5% in economics, 37.1% in physics), and we see that especially the Dutch cardiologists were less likely to reply, only 12.8% of them replied (see Table 1).[10] These biases were controlled with weighs in the bivariate analyses, see Section 3.4.

3.2 Dependent variables in the analyses

In the survey, the respondents were asked why they considered something to be the best research in their field and what was important for their assessments of grant proposals and candidates for academic positions. The two latter questions were only posed to respondents who indicated that they had conducted such reviews in the last 12 months.[11] Reply categories included various qualitative aspects and characteristics of good research as well as bibliometric indicators and open category answers (see survey questions in Supplementary Appendix).

The two kinds of assessments analysed in this paper—review of grant proposals and of candidates for academic positions—are performed in different settings. Research funding agencies and universities typically provide the contexts for these assessments. Within both types of organizations, the reviewers are normally provided with guidelines outlining the criteria and procedures for the review and are asked to compose a written review explaining their conclusions. Both types of assessments often include panel meetings in which the reviewers conclude on the ratings and/or ranking of the candidates/proposals.

The concerns and relevance of metrics in the reviews may vary greatly. When reviewing candidates for positions in research organizations, the reviewers are involved in facilitating or impeding the career of someone who might be their future colleague, and they often decide the composition of competencies and research interests at their own—or at a collaborating institution. This work may involve the reading and assessment of a considerable number of candidates' publications or simply assessing the publication lists based on metrics. Reviewer tasks for funding agencies may vary from assessing proposals for small individual grants to assessing those for long-term funding for large groups/centres, and from a few proposals close to their own field of research to many proposals assigned to a multi-disciplinary group of reviewers. The proposals may address a specific thematic call or a call open to all research questions, and the applicants' project descriptions and competencies are to be assessed accordingly.

In the survey we asked respondents about what they emphasized the last time they reviewed grant proposals, and what they emphasized the last time they reviewed a candidate for a position. They were also asked to indicate the type of grant/position in question. Metrics may be perceived as being less relevant as a basis for assessing junior applicants, i.e. applicants with a more limited track record. Hence, in this analysis, we distinguished between different types of positions and grants: recruiting to a junior or senior position; reviewing proposals for a research project, fellowship or large grant/centre, either to open calls or to targeted calls.

Table 1. Response rates by field and country

Country	Cardiology		Economics		Physics		Total	
	% replied	n	% replied	n	% replied	n	% replied	n
Netherlands	12.8	725	20.9	745	24.3	1010	19.9	2480
Norway	47.4	378	52.2	224	49.0	433	49.1	1035
Sweden	27.8	601	42.0	305	42.3	1526	38.7	2432
Total	25.8	1704	31.5	1274	37.1	2969	32.7	5947

3.3 Control variables

Research quality notions and assessments may differ between fields and countries, and may be influenced by the respondent's age, gender and academic position. Hence, in the analyses, we controlled for these variables, as well as for the type of grant or academic position being assessed. Table 2 provides details on the control variables. Note that all three fields studied are male-dominated. Even if the response rate among the female respondents was somewhat higher than among the male respondents, the obtained sample consists of 23% female respondents and 77% male respondents.[12]

In addition, we examined the relation between the respondents' publication outputs and their replies. The data on the respondents' publication output was collected from the Web of Science database (WoS) covering the 2011–2017 period, and included articles, reviews and letters published in journals indexed in WoS.[13] Three types of indicators were calculated. First, the number of publications per respondent during the period. Second, their mean normalized citation score (MNCS). Here, the citation numbers of each publication were normalized by subject field, article type and year, and then averages were calculated for the total publication output of each respondent. Third, their mean normalized journal score (MNJS) was determined, which involved similar calculations for the journals. The latter indicator is an expression of the average normalized citation impact of the journals in which the respondents have published their work, and high scores indicate that the respondents have published in a high-impact journal. On both indicators, 1.00

Table 2. Descriptive statistics

Variable/value	Count value	% value	n[a]
Age: 39 and younger	404	28	1435
Age: 40 to 49 years old	369	26	1435
Age: 50 to 59 years old	302	21	1435
Age: 60 years and older	360	25	1435
Gender: Female	325	23	1432
Gender: Male	1107	77	1432
Position: Assistant Professor	463	29	1611
Position: Associate Professor	391	24	1611
Position: Leader	77	5	1611
Position: Other	195	12	1611
Position: Professor	485	30	1611
Recruiting Juniors	552	71	774
Recruiting Seniors	222	29	774
Grant specification: Open Call	450	70	639
Grant specification: Target Research Call	189	30	639
Grant type: Fellowship	83	13	643
Grant type: Large Grants/Centre	78	12	643
Grant type: Research Project	482	75	643

Respondents' bibliometric performance	Mean	St. Dev.	Min	Max	n
Number of publications	27.86	63.835	0.250	781.00	1355
Log of number of publications	2.17	1.598	−1.39	6.66	1355
Have cited publications (dummy MNCS)	0.96	0.202	0.00	1.00	1355
MNCS	1.46	2.262	0.00	30.84	1355
Log of MNCS[b]	−0.03	0.913	−2.30	3.43	1297
MNJS	1.34	0.985	0.10	18.88	1355
Log of MNJS	0.15	0.517	−2.30	2.94	1355
Having publications in top percentile (dummy)	0.61	0.488	0.00	1.00	1355
Share of publications in top percentile	13.94	20.796	0.00	100.00	1355
Log of share of publications in top percentile[b]	2.73	0.933	0.00	4.61	828

[a]Smaller n on reviews of grant proposals and candidates for positions, as these questions were posed only to those who reported to have participated in such reviews the last 12 months.

[b]The log of MNCS/Share of publication in top percentile (including those who have scores above 0 on the MNCS indicator/publications in top percentile).

corresponds to the world average. As an additional citation indicator, the proportion of articles that are among the 10% most cited articles in their fields has been included (the share of publications in the top percentile can be found in Table 2).

We included these metrics in binary logistic regression analyses, investigating the relation between the respondents' bibliometric performance and their emphases on metrics when characterizing the best research in their fields, assessing grant proposals and assessing candidates for positions. Equations are attached in the note.[14] Apart from the factors included in the model, respondents' institutional affiliation and their specific research fields may influence their emphases in the assessment of research. Institutional affiliation has proven to influence researcher's evaluation at least in recruitments (Musselin 2010) and there may be large differences within research fields regarding notions of research quality and use of metrics (Lamont 2009; Hylmö 2018). Due to low numbers of respondents per institution, and insufficient data on subfields, we have not been able to control for these factors.

The bibliometric variables were skewedly distributed among the respondents, and thus the binary logistic regression analyses were conducted with log-transformed bibliometric variables, which ANOVA and AIC-tests showed improved our models. We settled on models with the log-transformed variables when displaying field differences. Still for graphic illustration of results the original variables are used to ease the interpretation for the reader. Table 2 displays the distribution of the original and log-transformed metrics variables.

It should be noted that the Web of Science database does not equally cover each field's publication output. Generally, physics and cardiology are very well encompassed, while the coverage of economics is somewhat less so, due to different publication practices (Aksnes and Sivertsen 2019). In addition, not all respondents had been active researchers during the entire 2011–2017 period, and for 16% of the respondents in the sample no publications were identified in the database. The latter individuals were not included in the bibliometric analyses. Despite these limitations, the data provides interesting information on the bibliometric performance of the researchers at an overall level.

3.4 Methods

We used the programming software 'R' when analysing the data and 'RMarkdown' for visualization. The RMarkdown file can be provided upon request.

Weighted Results: As sample sizes vary by fields and country, the bivariate analyses were weighted so that each field in each country

Table 3. Weights

Field	Sweden	Norway	The Netherlands
Cardiology	1.453	1.185	2.771
Economics	1.476	1.715	1.354
Physics	0.332	1.059	0.866

contributed equally to the totals (the weights are presented in Table 3). In the regression analyses, both field and country were controlled for, and the weights were not applied.[15]

Analyses: Binary logistic regression models were applied, including the stated characteristics of the best research, the emphasizes when assessing grant proposals and the emphasizes when assessing candidates for positions as dependent variables, while respondent characteristics (field, country, gender, age, academic position, and bibliometric performance) and type of proposal/position under review were included as control variables. To estimate whether the independent variables contributed with significant explanation to the variation in the dependent variable, we applied ANOVA tests (Agresti 2013). We further conducted AIC- and BIC-test to detect which independent variables best explained the independent variables (Agresti 2013) and applied the variance inflation factors-test (VIF-test) to check for eventual multicollinearity (Lin 2008). Finally, we checked for interaction effects between the independent variables. In the analyses, we used Sweden and economics as baseline categories; Sweden because it was the largest group and economics because it eased the interpretation of field differences (economics was the most deviant category). We also conducted the analyses with the other countries and fields as baseline categories to validate the presented results.[16]

We display the results from the binary logistic regression analyses in dot-and-whiskers plots with the fields' logit coefficients. In the graphs, economics is the baseline category (dotted line), and the likelihood of belonging to physics or cardiology is marked with standard errors. Hence, the graphs do not show potentially significant differences between physics and cardiology. In the (rare) cases wherein these differences are significant, this is commented on in the text. In addition, we illustrate results by calculating changes in probabilities on the dependent variables produced by the independent variables for selected subgroups. The full regression models are in the Supplementary Appendix Tables A1–A11.

4. Analysis: Metrics in peer assessments

4.1 Characteristics of the best research

As characteristics of the best research in one's field, impact factors and citations were among the less important aspects. In total, 22% of the respondents indicated journal impact factor and/or citation rates as reasons for considering something the best research in their field[17], whereas the most frequent reasons were that the research had solved key questions in their field (67%, see Table 4). Notably, respondents could select multiple replies and very few selected journal impact factor and/or citations as their only reasons for considering any research as being of the best.[18]

The binary logistic regression analysis indicates field-dependent reasons for considering something to be 'the best research', as illustrated in Figure 1 (see Supplementary Appendix Tables A1–A3 for full regression models). Economists were significantly more inclined than physicists to indicate journal impact factor as a reason for considering something the best research, but differences between the economists and cardiologists, or between the cardiologist and physicists, were not statistically significant. Moreover, the economists were more inclined than both the physicists and cardiologists to indicate many citations as a reason for considering any research to be the best. Interpreting the results, the regression coefficients imply that, for Swedish economists, the probability of answering high impact factor was 18%, while the probability for Swedish physicists was 14%. Similarly, the probability for Swedish economists to answer citations impact was 18%, while it was 10% for cardiologists and 13% for physicists in the same country.

Table 4. Reasons for considering something the best research in their field (Percent. Weighted results)

Reply	Cardiology	Economics	Physics	Total
Has answered/solved key questions/challenges in the field	70	62	70	67
Has changed the way research is done in the field (e.g. methodological breakthrough)	35	57	47	47
Has changed the key theoretical framework of the field	31	32	38	33
Has been a centre of discussion in the research field	30	29	34	31
Has benefitted society (e.g. appl. in industry, new clinical practices, informed public policy)	38	26	18	27
Has enabled researchers in the field to produce more reliable or precise research results	21	24	25	23
Was published in a journal with a high impact factor	18	21	13	17
Has attracted many citations	11	20	14	15
Has drawn much attention in the larger society	14	13	11	12
Is what all students/prospective researchers need to read	4	10	7	7
Other, please specify	2	2	2	2
Cannot say	2	3	1	2
n	405.25	405.25	405.25	1621

The ANOVA-analyses revealed country-dependent replies, but no dependence on the other control variables appeared (Supplementary Appendix Tables A1–A3). Interestingly, respondents in Norway were more inclined to indicate metrics as reasons for considering something to be the best research. Hence, country-related differences in adherence to metrics should be further explored.

4.2 Grant proposals

Whereas quantitative indicators appeared to have moderate importance in the identification of the best research in the field, 45% of those who had reviewed grant proposals replied that the *number of publications/productivity* was 'highly important' in their assessments of the best proposal, and 23% found *citations* 'highly important' in their assessments. These metrics were also relatively

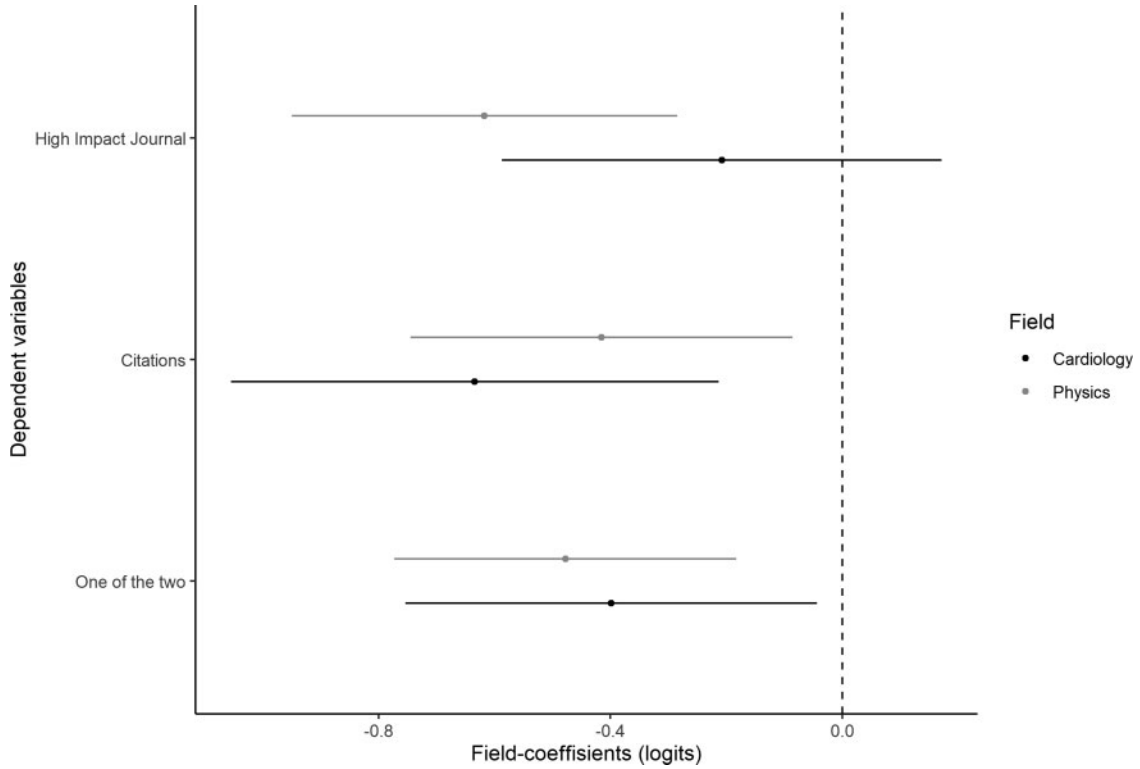

Figure 1. Journal impact factor and citations as reasons for considering something the best research in the field. Field coefficients from binary logistic regression analyses (Dot-And-Whiskers Plots). Economics as the baseline category represented by the dotted line.

Table 5. Aspects identified as "highly important" in grant assessments (Percent. Weighted results).

Reply	Cardiology	Economics	Physics	Total	Total n
Project description: research question/problem selection	98	87	94	94	678
Project description: methods/research plan	90	82	81	85	670
Track record of the research team: important prior contributions in the relevant research field (assessed independently of citation scores and source of publication)	46	35	55	46	673
Track record of the research team: number of publications/productivity	41	50	44	45	674
The research environment: resources and facilities for performing the proposed research	59	15	41	41	671
Track record of the research team: citation impact of past publications	18	29	25	23	676
Track record of the research team: experience with risk-taking research	19	14	21	19	670
Communication / dissemination plan for scientific publications	9	10	9	9	673
Other, please specify	10	8	9	9	664
Communication/dissemination plan addressing user groups outside academia	6	6	5	6	666

important compared to several other aspects (Table 5). They still appear far below the 'research question' (94%) and the 'methods/research plan' (85%), which came up as the most important in the assessments. Nonetheless, including those who replied 'somewhat important' (48% for number of publications and 59% for citation impact), the great majority replied that such metrics impacted their assessments of which proposal was the best (Supplementary Appendix Table A12).

The binary logistic regression analysis shows that emphases on metrics were field-dependent, as shown in Figure 2 (full regression models are shown in Supplementary Appendix Tables A4–A7). Compared to cardiologists, the economists (dotted line) were significantly more inclined to identify the number of publications and citation impact as 'highly important'. Conversely, the physicists were significantly more inclined to emphasize important research contributions (assessed independently of metrics) than were the

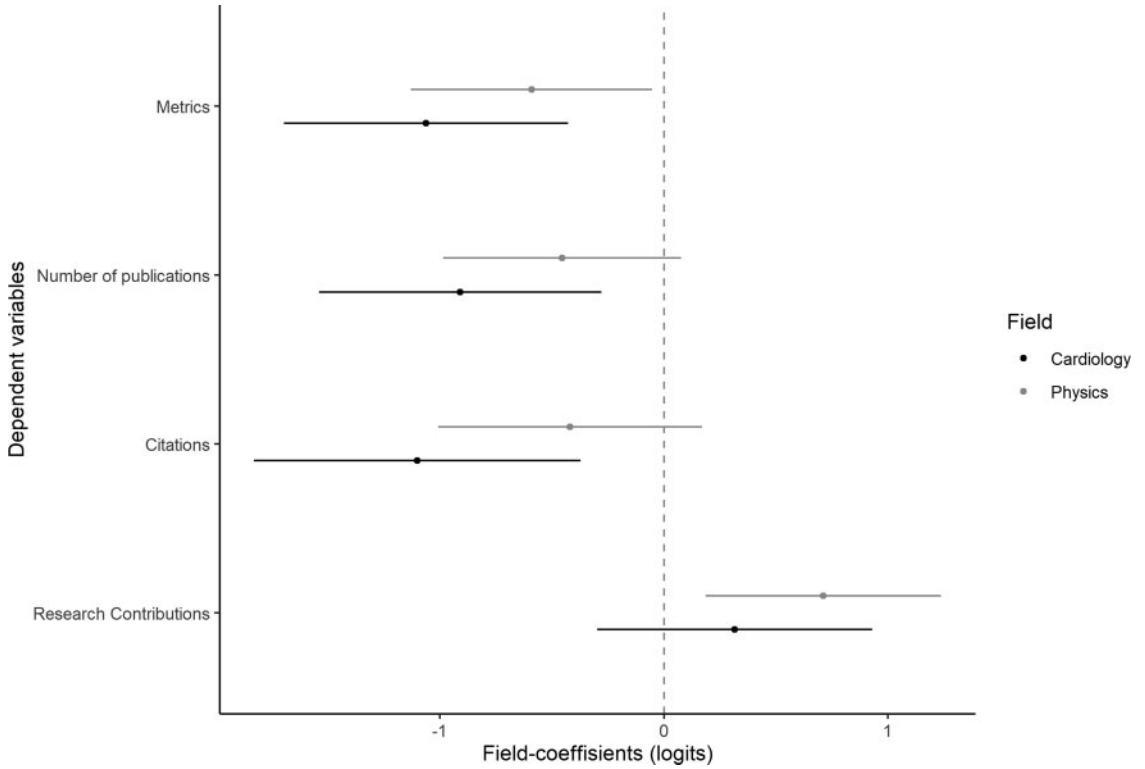

Figure 2. Aspects identified as 'highly important' in assessing grant proposals. Field coefficients from binary logistic regression analyses (Dot-And-Whiskers Plots). Economics appears as the baseline category, represented by the dotted line.

economists. The analysis uncovered less difference between physicists and cardiologists, but still, the physicists were significantly more inclined than the cardiologists to emphasize citations.

The regression coefficients imply that the probability of Swedish economist professors with an average share of top publications and number of publications, of identifying the number of publications or/and citations as 'highly important' in their assessment of proposals to open calls is 50%, whereas the probability of the similar groups of physicists and cardiologists to do so is substantially lower (39% for physicists and 29% for cardiologists). Conversely, the economists in this group (professors with average bibliometric scores) were less inclined to emphasize 'important prior research contributions assessed independently of metrics' (59% for physicists, 49% for cardiologists, and 42% for economists).

Furthermore, the regression analyses indicated insignificant country-related effects, but significant effects of the respondents' academic positions and the type of grants being reviewed. The probability of identifying citations or number of publications as highly important was lower when assessing project grants than when assessing fellowships or large/centres grants. Moreover, the probability of identifying the number of publications as highly important was lower when reviewing proposals to open calls rather than targeted calls (Supplementary Appendix Tables A4–A7). The replies also depended on the respondents' bibliometric performance, as discussed in detail below.

4.3 Candidates for positions

Similar results to those for assessing grant proposals appear for the assessment of candidates for positions: quantitative measures appear more important than when identifying the best research in the field. Forty-two percent answered that the number of publications/productivity was 'highly important' in their assessments of candidates (Table 6). Citations impacts appear to be less important (19% replied that this was highly important). Notably, research contributions assessed independently of citation scores and publication source appear more important than number of publications/productivity in cardiology (47% highly important) and physics (61% highly important). In economics, on the other hand, there is a higher percentage who find the number of publications/productivity to be highly important (54%) and a lower percentage who find that contributions assessed independently of metrics are highly important (45%).

When the respondents were asked to identify the most important among the aspects they had identified as highly important, the candidate's 'potential for future achievements' and 'expertize matching the group/unit/project' prevail as the two most important aspects in all three fields, indicating that these have general high importance regardless of fields. The third most important aspect, however, varied greatly between the fields: whereas cardiology appears with 'general impression from interview with candidate' and physics with 'important prior research contributions (assessed independently of citation scores and source of publication)', in economics 'number of publications/productivity' appears as the third most important aspect (Supplementary Appendix Figure A1).

Binary logistic regression analysis confirms statistically significant differences between fields (documented in Supplementary Appendix Tables A8–A11 and illustrated in Figure 3 below). Economists were more inclined than both cardiologists and physicists to identify the number of publications as 'highly important' when assessing candidates for positions (the difference between the

Table 6. Aspects identified as 'highly important' when assessing candidates for positions (Percent. Weighted results).

Reply	Cardiology	Economics	Physics	Total	Total n
The potential for future achievements	83	90	87	87	823
Matching field/expertise to the needs of the group/unit/project	70	65	72	69	816
General impression from interview with candidate	74	52	69	65	813
Communication and language skills	61	40	53	51	821
Research achievements: important prior research contributions (assessed independently of citation scores and source of publication)	47	45	61	51	810
Research achievements: number of publications/productivity	34	54	39	42	817
Ability to compete for research grants	43	21	33	32	814
Standing of the unit/group where the candidate is/has been working/trained	24	20	16	20	815
Research achievements: citation impact of past publications	17	19	21	19	808
Teaching experience/achievements (including supervision of students)	18	18	16	18	816
Ensure diversity in the group/department (e.g. gender, ethnicity, age)	13	11	17	14	816
Other, please specify	12	5	23	13	95
Experience/achievements from work outside science	10	4	4	6	815
Experience in interacting with the public/users/industry	8	2	3	4	805

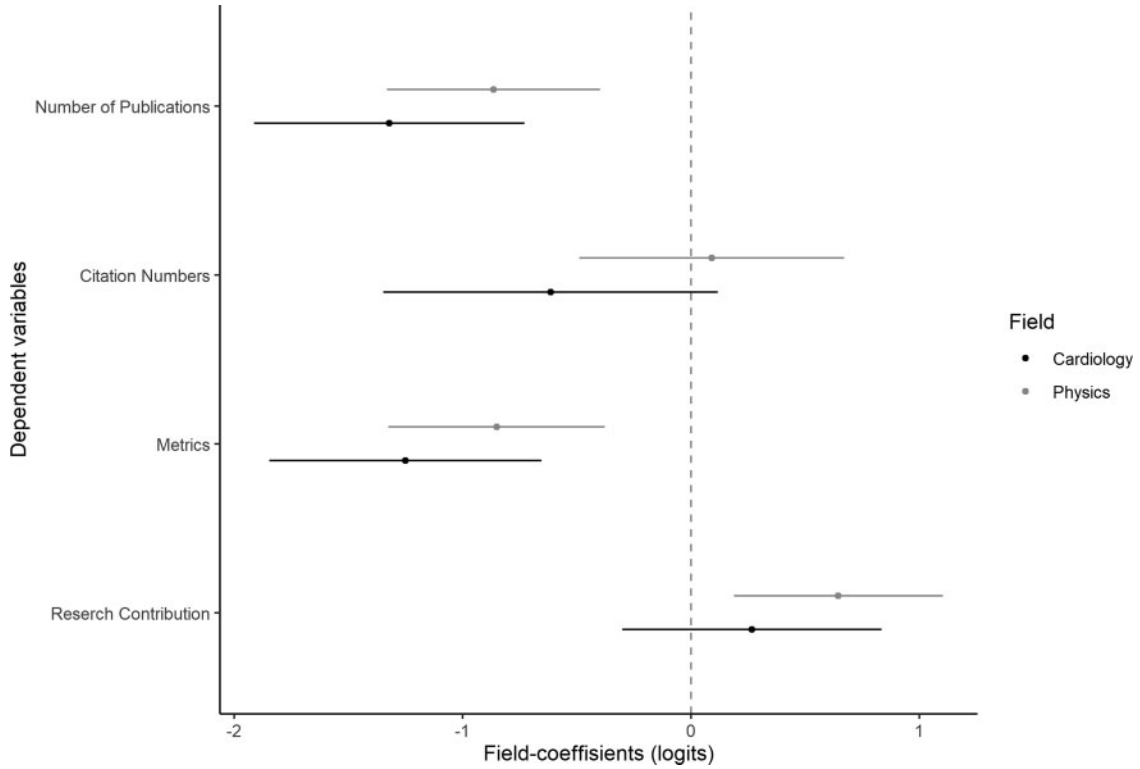

Figure 3. Aspects identified as an 'highly important' when assessing candidates for positions. Field coefficients from binary logistic regression analyses (Dot-And-Whiskers Plots). Economics as the baseline category represented by the dotted line.

two latter fields was not significant). Moreover, physicists, more frequently than economists, answered that prior research contributions had been 'highly important'. The regression coefficients for a reference group of Swedish professors with average scores on the bibliometric indicators (number of publications and share of top cited publications) who assess candidates for senior positions, show that the probability of stating that the 'number of publications' is 'highly important' was 83% in economics, 68% in physics and 57% in cardiology. In contrast, the probability in this group of answering that prior contributions were 'highly important' was 84% in physics, 76% in cardiology, and 72% in economics.

In sum, the results indicate some field differences in line with the different publication and authorship patterns noted in Section 2. Economics, the field with the lower average number of publications per author and lower average number of co-authors, relies more frequently on number of publications/productivity when assessing candidates for positions. Conversely, higher numbers of co-authors and publications appear in physics and cardiology compared to economics (Piro, Aksnes and Rorstad 2013), and this may be a reason for less emphasis on number of publications in the former fields. In these fields, it may be far less straightforward to reach conclusions based on the length of individual researchers' publication lists.

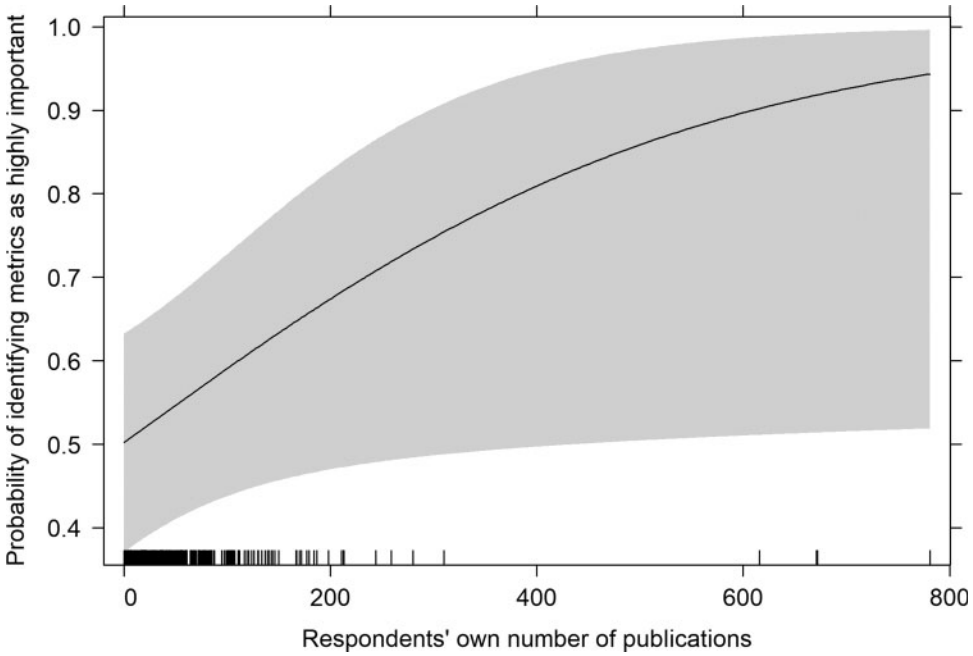

Figure 4. Assessment of grant proposals: The probability of identifying the number of publications and/or citation impact as 'highly important', by respondents' number of publications. The rug at the x-axis marks the number of observations.

The ANOVA tests showed that, in some of the models, the respondents' gender, age and country had significant effect on the respondents' emphases. The country-related effects were mostly insignificant, but being Dutch instead of Swedish decreased the possibility of identifying publication numbers as highly important. Professors were less inclined than those in other kinds of positions to see the number of publications as 'highly important', while there was no significant effect on emphasis on citations. Furthermore, both quantitative measures and important prior contributions were more often seen as important when recruiting to senior rather than to junior positions.

4.4 The effects of the respondents' bibliometric performance

Looking further into the results, a key question is whether the respondents' emphases on publication metrics corresponded with their own bibliometric performance. For example, one might find that researchers with many publications would put more emphasis on this dimension in their assessments. Therefore, we compared the respondents' answers with their own scores on the relevant bibliometric indicators.

The regression analyses showed no effect of respondents' bibliometric performance on their reasons for considering something the best research. However, in practice, when assessing grants proposals and candidates for positions, their own performance was positively related to their use of metrics. When assessing grant proposals, the probability of identifying the number of publications and citation impact as 'highly important' increased along with the respondents' number of publications, whether they had top percentile publications, and their share of top percentile publications.[19] Figure 4 displays this relationship, and shows how the probability of identifying citation impact and/or number of publications as highly important in assessments of grant proposals for Swedish professors in economics increases with their own number of publications.[20] Respondents with high bibliometric performance scores more frequently considered such indicators as important in their assessments. On the other hand, the respondents' bibliometric performance did not affect whether they found prior contributions 'highly important' (Supplementary Appendix Table A7).

Similarly, metrics in the assessment of candidates for academic positions depended on the respondents' bibliometric performance, but less so than for grant assessments. The respondent's own number of publications did not significantly affect the probability of identifying candidates' citation impact or number of publications as 'highly important', yet the log of the number of publications increased with respondents' publication output. Moreover, a respondent having top cited publications increased the probability of identifying candidates' citations as 'highly important'. For Swedish economics professors who recruited for senior positions, having top cited publications increased this probability from 28 to 40%. However, neither the respondents' MNCS, MNJS nor share of top cited publications had a significant effect on the use of metrics in the assessment of the candidates. Hence, the importance of metrics in these assessments was lower than that for the assessment of grant proposals.

Moreover, the respondents' MNCS (log of) and MNJS increased the probability of identifying prior research contributions as highly important when assessing candidates for positions, but did not increase the probability of identifying such contributions as highly important when assessing grant proposals.

4.5 Divergent opinions and perspectives

Insofar as this attempt to conclude on whether metrics are considered a legitimate and integral part of assessments of research, the results indicate conflicting views as well as differences between review contexts and the type of metrics. A large majority of

Table 7. Free text replies—what was important for your assessment of:

The best grant proposal	
1	'I evaluate research according to the value of proposed research. For the highest scores, there has to be an outstanding problem to address. There has to be a realistic plan outlining how this is possible. [. . .] I don't pay much attention to the rate of publications, but rather to the lasting impact of these. I also do not care much about citations, as this varies profoundly between different topics. Rather, impact must be assessed based on an actual understanding of previous research' (Research project grant, physicist in Sweden who selected 'not important' both on number of publications and citations.)
2	'Solves relevant questions. Science at excellent level considering modern perspectives in research evaluation (NOT publication and citations numbers as primary component).' (Fellowship grant, cardiologist in Norway who selected 'not important' both on number of publications and citations)
3	'Concrete yet ambitious proposal, novel methodology, included a Plan B if the risky plan A failed, collaboration with industry to obtain interesting field data, productive researcher with high H-index.' (Research project grant, social sciences/other in the Netherlands)
4	'Important research problem, High-quality candidate(s), Track record in terms of publications in top international journals (e.g. Nature, Science), assessed independently of citation records.' (Large grant/centre, physicist in the Netherlands)
The best candidate for a position	
5	'The candidate was expected to do research, to teach, and in particular to build a Research Group within [subfield]. Communication skills, Cultural competence and networking ability are crucial, in addition to number and "weight" of publications.' (Junior/early career position, cardiologist in Norway)
6	'Much of the selection is based on the impact factors of the journals a candidate has published in, and potentially the network of the candidate (I do not necessarily believe these are the best criteria per se, but they are generally used).' (Junior/early career position, economist in the Netherlands)
7	'It's a mix. Citation impact without productivity indicates a few very highly cited papers, which is not what I mean. It should be a combination of high productivity of high-quality papers that also have attracted citations. So number of high-quality papers, overall citations, h-index, and where the work was published all matter.' (Senior/tenure position, physicist in Sweden)

respondents confer to metrics in their reviews and seem to find it a legitimate and ordinary basis for reviews. This goes particularly for the number of publications/productivity in the review of grant proposals; only six percent replied that this was not important (Supplementary Appendix Table A12). Still, a substantial proportion (33%) indicated that citation scores were not important regarding assessments of candidates for academic positions (Supplementary Appendix Table A13).

The free text replies concerning the main positive characteristics of the best proposals illustrate the divergent opinions and perspectives. Some grant reviewers emphasized that metrics were not important (illustrated by #1 and #2 in Table 7). Others emphasized publication and citation rates as key characteristics of the best proposal, or simply publications in top international journals (#3 and #4 in Table 7).

Several of those who had reviewed candidates for positions seemed to find publications in major/top journals a basic or objective criterion, and then added other important characteristics that would befit the particular research group or the tasks of the position (#5 in Table 7). Others indicated that the ranks of the journals the candidate had published in—or a combination of relevant metrics—was important in the selection process. Still, views on the adequacy of such criteria varied (#6 and #7 in Table 7).

5. Discussion and implications

In this paper we have explored whether metrics are part of researchers' notion of good research, and whether metrics are used when reviewing research. Concerning the first issue, only a minority of the respondents reported metrics as a reason for considering something to be the best research. Thus, the empirical support for such an association is generally weak. On the second question, we find strong supportive evidence as a large majority indicated that metrics were important or partly important in their review of grant proposals and assessments of candidates for academic positions.

Notably, drawing conclusions on researchers' notions of research quality is difficult. Research quality is a multidimensional concept; what are seen as the key characteristics of good research may differ largely between contexts and fields (Langfeldt et al. 2020). Metrics such as citations, publication counts or journal impact factors may be perceived as relating to different characteristics of research quality, e.g. according to bibliometric studies, citations reflect (to some extent) the scientific value and impact of research, but not its originality, plausibility/soundness or societal value (Aksnes, Langfeldt and Wouters 2019). Our data indicate that the respondents distinguish between quantitative indicators as proxies for success when assessing the potential of future projects or candidates for positions and what they hold to be the characteristics of good research. A large majority of the respondents reported metrics as highly or somewhat important in their reviews of grant proposals and of candidates for positions, whereas about one-fifth of them indicated that one of their reasons for concluding on what was the best research in their field was that it was published in a journal with high impact factor or that it had attracted many citations. Hence, for one-fifth of the researchers in the survey, metrics seem to be a judgement device when identifying good research within one's

own field. This does not necessarily imply that they hold metrics, as such, to be characteristics of good research. Very few respondents indicated the journal impact factor or high citation rates as sole indicators of the best research in their field, and there is little indication that respondents view quantitative indicators as being a sufficient basis for concluding on eminent science. Nevertheless, some have suggested that publishing in high-impact journals has become an independent measure of scientific quality (Wouters 1999; Rushforth and de Rijcke 2015).

Moreover, the analysis indicates significant field differences in the use of publication metrics: The economists were more inclined to indicate journal impact factor and many citations as reasons for concluding that something is the best research in their field, and they were more inclined to emphasize the applicants' number of publications when assessing grant proposals and candidates for positions. Physicists and cardiologists, on the other hand, were less inclined to emphasize metrics and more inclined to emphasize prior research contributions assessed independently of metrics. These differences go along with differences in how research is organized and valued in these fields. In economics, research is mostly performed by individuals and organized around a theoretical core and key journals of high importance for individual reputation (Whitley 1984; Hammarfelt and Rushforth 2017; Hylmö 2018). Herein, high reliance on metrics may be explained by the combination of an explicit journal hierarchy and organization of research that makes it easier to attribute research performance to individuals. Physics consists of highly collaborative fields, some with hyper-authorship (Birnholtz 2008), and using publication metrics to attribute research performance to individual researchers is more difficult. Similarly, cardiology is a field within medical research with specialized tasks and skills, highly dependent on collaboration, resources and facilities for performing research (Whitley 1984), which may explain the lower emphasis on publication metrics than in economics, as well as far stronger emphases on research resources and facilities when assessing grant proposals. Notably, there is also much variation in replies within the fields: for example, a substantial proportion of the physicists and the cardiologists indicate the applicant's number of publications as highly important when assessing grant proposals and candidates for academic positions, whereas others find it unimportant or somewhat important. In sum, this points to the importance of understanding how epistemic and organizational differences—both between and within research fields—generate different bases for assessing research and research performance, and thereby different use of metrics.

Despite our comparative point of departure, along with the inclusion of countries with different use of metrics in national research funding, we found only limited country-specific differences. The lack of country-related differences indicates that notions of research quality are more connected to general field differences than to national context (Lamont 2009; Musselin 2010). Still, even if our sample of three countries in the northern corner of Europe represents variety in research funding and research evaluation, a larger sample of more diverse countries might have exposed greater differences in the use of metrices in peer assessments.

The findings have policy importance for multiple aspects of the evaluation of research. Below, we discuss implications relating to (1) how research agendas and research activity adapt to research evaluations, (2) the policies for restraining the (mis)use of metrics in research evaluation, and (3) the design and organization of research evaluations.

First, an emphasis on metrics may impact research activity and research agendas. Researchers—at least young and non-tenured ones—cannot disregard what gives acclaim in the academic career system and what is needed for attracting research funding. They need to take into consideration what kind of research will help them qualify for grants and positions (Müller and de Rijcke 2017). Notably, in our data, economists seem to put less emphasis (compared with the other groups) on expertize, matching the needs of the research group/unit, and they seem to be more apt to use metrics (Supplementary Appendix Figure A1). This may imply that, rather than making explicit decisions about the kind of researchers to employ (their topics and methods), the researchers who are able to do the kind of research that are most easily published in (top) economics journals are hired (Lee, Pham and Gu 2013). Hence, the ways in which researchers adapt to metrics come up as a key topic for studies in research evaluation and, more generally, for research policy.

Second, despite increasing concerns in the scientific communities on the use and misuse of research metrics (Wilsdon et al. 2015), the results herein indicate that researchers rely on the three types of metrics addressed in the survey: journal impact factors, number of publications and citation impact. Close to one-fifth of the respondents reported high impact factor as a reason for something being the best research in their field. As discussed in the introduction, journal impact factors and journal rankings have been widely used, particularly in medicine and economics, for assessing scientific performance. With the launch of the DORA-declaration in 2012, the problem with this practice has received more attention.[21] As a response, policies and practices of many funding organizations, scientific societies, institutions and journal publishers have changed, according to Schmid (2017). Nevertheless, others report that journal impact factors are still used for purposes that conflict with the DORA-declaration (Bonnell2016). Notably, the DORA-declaration has led to an increased focus on other ways to assess research. This includes the development of alternative paper-based metrics (Schmid 2017). Indicators of number of publications and citation impact do not have similar problems to those associated with the journal impact factor. Nevertheless, it is well known that these indicators also have various limitations and shortcomings as performance measures, particularly when applied at micro levels (Wildgaard, Schneider and Larsen 2014), and our survey indicates extensive use of these indicators at micro levels when reviewing grant proposals or candidates for academic positions. Moreover, the field differences found in the survey point to a need for a better understanding of why and how metrics are used in different fields as well as a need to consider field-adjusted policies for the use of metrics in research evaluation.

Finally, there are implications regarding the design and organization of research evaluation. Publication-based metrics seem to be perceived as good proxies for research quality and performance, at least for the majority of the researchers in the fields studied. This may be because they trust the review processes of the scholarly journals and publishers in their field, and metrics make sense as a proxy for quality. From this perspective, the editors and reviewers of the major journals end up high on the list of those controlling the gatekeeping criteria, not only for scientific publishing, but also for academic positions and research grants. At the end of the 'review chain', we will often find the criteria, review processes and publication policy of the major journals in the field. Hence, the researchers complying with the topics, perspectives/methods and formats of these journals can be expected to have the highest chances of success in competitions for grants and academic positions. Still, the above

analysis indicates deviant views among reviewers on the use of metrics in research evaluation. So even if certain topics, perspectives or methods dominate a field, the outcome of review processes may vary by the panel members' views on metrics. Consequently, when it comes to the 'luck of the reviewer draw' (Cole, Cole and Simon 1981), not only the panel members' scholarly profile and competences, but also their preferences for metrics may be decisive. This implies that in order to provide fair and well-grounded review processes, there is a need for insight into how panels use metrics in their assessments and to encourage explicit discussions about the use of metrics[22]. If the role of metrics is not openly discussed in review panels, nor understood by those organizing the reviews and acting upon them, we risk concealed review criteria.

Notes

1. When applying for advanced grants from the European Research Council (ERC), applicants have been asked to provide a ten-year track record including publications in leading journals and 'indicating the number of citations (excluding self-citations) they have attracted (if applicable)'. https://erc.europa.eu/sites/default/files/document/file/ERC_Work_Programme_2015.pdf. We find this formulation in the ERC work programmes for 2008–2016. For 2017, 2018, and 2019 the wording is: '(properly referenced, field relevant bibliometric indicators may also be included)'. http://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/erc/h2020-wp19-erc_en.pdf.
2. http://www.ascb.org/dora/
3. Likewise, in Norway, the research council regularly conducts peer evaluations of disciplines and subjects as well as institutes and programmes, and bibliometric indicators are used as one source of information whenever relevant (Sivertsen 2017).
4. Such perceptions may in turn be formed by/rooted in extensive use of e.g. journal rankings or citation measures in the field (Espeland and Sauder 2007: 16).
5. The use of metrics in peer review is also part of the more general story about how information technology impacts our evaluative practices (Lamont 2012: last section).
6. For example, the Research Council of Norway requires applicants to use a CV template that includes citation counts for applicants for regular researcher projects in all research fields. Up to 2018, the RCN template was named after its role model 'ERC track record description'.
7. A majority of those who provided input to the 'Metric Tide' report were sceptical to the role of metrics in research management while a significant minority were more supportive of the use of metrics (Wilsdon et al. 2015: viii).
8. The minimum number of publications (5 for economics and 10 for cardiology and physics) was selected based on analyses of individual publication output during the 2011–2016 period. By applying these thresholds we aimed at including the more active researchers within the fields and leaving peripheral researchers out. A higher number was applied for cardiology and physics because of the higher publication frequencies (and co-authorship) in these fields.
9. The survey is part of a larger research project and was launched in five countries in 2017–2018. The present analysis is based on replies from higher education institutions in three of these countries (1,621 replies). The full survey included 2,587 replies, and is also comprised of replies from economics and physics in Denmark and the UK, as well as replies from researchers affiliated with independent research institutes. Replies from Denmark and the UK are excluded from the present analyses, as cardiology was not sampled in these countries. We checked for the impacts of excluding the UK and Danish samples by conducting the analyses on Economics and Physics in all five countries, and did not find any significantly deviant results. Moreover, replies from independent research institutes are excluded as they constitute a small sample (in total 111 replies in the three countries) and research settings which may differ substantially from those at higher education institutions.
10. Table 1 shows response rates by field as identified in the sampling process, whereas our analyses are based on field as identified by survey responses. Respondents who replied other fields of research, rather than 'Cardiac/cardiovascular systems/diseases', 'Economic' or 'Physics' are not included in the analysis. Hence, the analyses are based on a smaller sample (1,621 respondents) than that which prevails in Table 1 (1,942 respondents).
11. Consequently, the analyses are based on the full sample for the first question, and different subsamples for the two latter questions. We checked for impacts of sample variation by additional analyses of those included in both subsamples (451 respondents stated that they had reviewed both grant proposals and candidates for positions the last 12 months). These analyses did not give deviant results. Hence, differences between the two review settings appearing from our data are not due to different subsamples.
12. We have data on gender for 92% of the invited respondents. Of these, 39% of the female and 35% of male respondents replied. Of those without information on gender, we have replies from 4%.
13. We have excluded minor contributions such as editorials, meeting abstracts, and corrections. As letters usually do not represent full scientific contributions, they are weighted as 0.25 of an article; this is in accordance with principles often applied by the Centre for Science and Technology Studies (CWTS) of Leiden University (for further discussion, see van Leeuwen, van der Wurff and de Craen 2007).
14. $Y1 = X1Country + X2Field + X3Bibliometricsb + X4Age + X5Gender + X6Position + e$
    $Y2 = X1Country + X2Field + X3Bibliometricsb + X4Age + X5Gender + X6Position + X7Call + X8Type + e$
    $Y3 = X1Country + X2Field + X3Bibliometricsb + X4Age + X5Gender + X6Position + X7Vacancy + e$
    Dependent variables: why they considered something to be the best research in their field (Y1), what was important for their assessment of grant proposals (Y2) and candidates for positions (Y3).
    a = type of assessment criteria
    b = type of bibliometrics
    e = error term
15. As an extra control, the regression analyses were run with weights. Results were not altered.
16. We also conducted ordinal logistic regressions for the best suited models with assessment of grant proposals and candidates for academic positions as dependent variables. These

models confirmed the results of the binary logistic models, with the exception that the respondents' share of top percentile publications did not have a significant effect on their emphases on numbers of publications when assessing grant proposals. Likewise, the respondents' fields of research did not have a significant effect on their emphases on citation impact when assessing grant proposals. Still, BIC-tests indicated that the binary logistic models were better suited to describe the data, and as these results are easier to communicate, we chose to keep them.

17. Of these, 17% replied high impact factor, 15% many citations (Table 4).
18. In total, 17 respondents selected journal impact factor and citations as the only reasons, five selected only journal impact factor and four selected only citations.
19. Their MNCS did not affect their use of metrics, but the log-transformed MNCS variable showed increased use of metrics with increasing (log of) MNCS (Supplementary Appendix Tables A4–A6).
20. As mentioned, the respondents' number of publications was very skewed. The black line at the x-axis (the rug) shows that most respondents had between 0 and 100 publications.
21. Here it was declared that journal-based metrics, such as journal impact factors should not be used as 'a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions' (http://www.ascb.org/dora/). Currently more than 1,800 organizations and 15,500 individuals have signed the declaration.
22. A study of grant panels at the UK National Institute for Health Research indicated that their panel members primarily use the metrics provided to them in their individual assessments in advance of the panel meeting, and less in the panel discussion (Gunashekar, Wooding and Guthrie2017).

Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

Acknowledgements

The research was funded by the Research Council of Norway, grant number 256223 (the R-QUEST centre). The multinational survey analysed in the paper was a joint effort of the R-QUEST team. Thed van Leeuwen took an important role in the sampling and in providing the bibliometric indicators. We are thankful to Thed van Leeuwen, Anders Hylmö, Thomas Franssen and the rest of the R-QUEST team for input and comments to the paper.

References

Aagaard, K. (2015) 'How Incentives Trickle down: Local Use of a National Bibliometric Indicator System', *Science and Public Policy*, 42: 725–37.

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., and Van Noorden, R. (2010) 'Do Metrics Matter?', *Nature*, 465: 860–2.

Agresti, A. (2013) *Categorical Data Analysis*, 3rd edn. New Jersey: John Wiley & Sons.

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019) 'Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories', *SAGE Open*, 9: 1–17.

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

Aksnes, D. W., and Sivertsen, G. (2019) 'A Criteria-Based Assessment of the Coverage of Scopus and Web of Science', *Journal of Data and Information Science*, 4: 1–21.

Allen, M. A. (2010) 'On the Current Obsession with Publication Statistics', *ScienceAsia*, 36: 1–5.

Ball, R. (2017) *An Introduction to Bibliometrics. New Developments and Trends*. Cambridge, MA: Chandos Publishing.

Birnholtz, J. (2008) 'When Authorship Isn't Enough: Lessons from CERN on the Implications of Formal and Informal Credit Attribution Mechanisms in Collaborative Research', *The Journal of Electronic Publishing*, 11:

Bollen, J., Rodriguez, M. A., and Van De Sompel, H. (2006) 'Journal Status', *Scientometrics*, 69: 669–87.

Bonnell, A. G. (2016) 'Tide or Tsunami? The Impact of Metrics on Scholarly Research', *Australian Universities' Review*, 58: 54–61.

Bornmann, L., Butz, A., and Wohlrabe, K. (2018) 'What Are the Top Five Journals in Economics? A New Meta-Ranking', *Applied Economics*, 50: 659–75.

Brown, H. (2007) 'How Impact Factors Changed Medical Publishing - and Science', *British Medical Journal*, 334: 561–4.

Coats, A. J. S., and Shewan, L. G. (2015) 'Impact Factor: Vagaries, Inconsistencies and Illogicalities; Should It Be Abandoned?', *International Journal of Cardiology*, 201: 454–6.

Cole, S., Cole, R., and Simon, G. A. (1981) 'Chance and Consensus in Peer Review', *Science*, 214: 881–6.

Cronin, B. (2001) 'Hyperauthorship: A Postmodern Perversion or Evidence of a Structural Shift in Scholarly Communication Practices?', *Journal of the American Society for Information Science and Technology*, 52: 558–69.

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

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., and Hammarfelt, B. (2016) 'Evaluation Practices and Effects of Indicator Use – A Literature Review', *Research Evaluation*, 25: 161–9.

Espeland, W. N., and Sauder, M. (2007) 'Rankings and Reactivity: How Public Measures Recreate Social Words', *American Journal of Sociology*, 113: 1–40.

Gibson, J., Anderson, D. L., and Tressler, J. (2014) 'Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California', *Economic Inquiry*, 52: 1322–40.

Glänzel, W., and Moed, H. F. (2002) 'Journal Impact Measures in Bibliometric Research', *Scientometrics*, 53: 171–93.

Gunashekar, S., Wooding, S., and Guthrie, S. (2017) 'How Do NIHR Peer Review Panels Use Bibliometric Information to Support Their Decisions?', *Scientometrics*, 112: 1813–35.

Haddow, G., and Hammarfelt, B. (2019) 'Quality, Impact, and Quantification: Indicators and Metrics Use by Social Scientists', *Journal of the Association for Information Science and Technology*, 70: 16–26.

Hammarfelt, B. (2018) 'Taking Comfort in Points: The Appeal of the Norwegian Model in Sweden', *Journal of Data and Information Science*, 3: 85–95.

Hammarfelt, B., and Rushforth, A. D. (2017) 'Indicators as Judgment Devices: An Empirical Study of Citizen Bibliometrics in Research Evaluation', *Research Evaluation*, 26: 169–80.

Heckman, J. J., and Moktan, S. (2018) *Publishing and Promotion in Economics: The Tyranny of the Top Five. NBER Working Paper No. 25093*. Institute for New Economic Thinking.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. (2015) 'The Leiden Manifesto for Research Metrics', *Nature*, 520: 429–31.

Hug, S., and Aeschbach, M. (2020) 'Criteria for Assessing Grant Applications: A Systematic Review', *Palgrave Communications*, 6: 1–15.

Hylmö, A. (2018) 'Disciplined Reasoning: Styles of Reasoning and the Mainstream-Heterodoxy Divide in Swedish Economics', Doctoral thesis, Lund University, Department of Sociology.

Jonkers, K., and Zacharewicz, T. (2016) *Research Performance Based Funding Systems: A Comparative Assessment*. Luxembourg: Publications Office of the European Union.

Kalaitzidakis, P., Mamuneas, T. P., and Stengos, T. (2011) 'An Updated Ranking of Academic Journals in Economics', *Canadian Journal of Economics-Revue Canadienne D Economique*, 44: 1525–38.

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

Lamont, M. (2012) 'Toward a Comparative Sociology of Valuation and Evaluation', *Annual Review of Sociology*, 38: 201–21.

Langfeldt, L., Nedeva, M., Sörlin, S., and Thomas, D. A. (2020) 'Co-Exiting Notions of Research Quality: A Framework to Study Context-Specific Understandings of Good Research', *Minerva*, 58: 115–37.

Langfeldt, L., and Scordato, L. (2016) *Efficiency and Flexibility in Research Funding. A Comparative Study of Funding Instruments and Review Criteria. NIFU Report 9/2016*. Oslo: NIFU Nordic Institute for Studies Innovation, Research and Education.

Lee, F. S., Pham, X., and Gu, G. (2013) 'The UK Research Assessment Exercise and the Narrowing of UK Economics', *Cambridge Journal of Economics*, 37: 693–717.

Lewison, G., Cottrell, R., and Dixon, D. (1999) 'Bibliometric Indicators to Assist the Peer Review Process in Grant Decisions', *Research Evaluation*, 8: 47–52.

Lin, F. (2008) 'Solving Multicollinearity in the Process of Fitting Regression Model Using the Nested Estimate Procedure', *Quality and Quantity*, 42: 417–26.

Loomba, R. S., and Anderson, R. H. (2018) 'Are we Allowing Impact Factor to Have Too Much Impact: The Need to Reassess the Process of Academic Advancement in Pediatric Cardiology?', *Congenital Heart Disease*, 13: 163–6.

Martin, B. R. (1996) 'The Use of Multiple Indicators in the Assessment of Basic Research', *Scientometrics*, 36: 343–62.

Moed, H. F. (2005) Citation Analysis in Research Evaluation. Dordrecht: Springer.

Müller, R., and de Rijcke, S. (2017) 'Thinking with Indicators. Exploring the Epistemic Impacts of Academic Performance Indicators in the Life Sciences', *Research Evaluation*, 26: 157–68.

Musselin, C. (2010) *The Market for Academics*, New York: Routledge.

Piro, F. N., Aksnes, D. W., and Rorstad, K. (2013) 'A Macro Analysis of Productivity Differences Across Fields: Challenges in the Measurement of Scientific Publishing', *Journal of the American Society for Information Science and Technology*, 64: 307–20.

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: 117–39.

Schmid, S. L. (2017) 'Five Years post-DORA: Promoting Best Practices for Research Assessment', *Molecular Biology of the Cell*, 28: 2941–4.

Sivertsen, G. (2017) 'Unique, but Still Best Practice? the Research Excellence Framework (REF) from an International Perspective', *Palgrave Communications*, 3: 17078.

Söderlind, J., and Geschwind, L. (2020) 'Disciplinary Differences in Academics' Perceptions of Performance Measurement at Nordic Universities', *Higher Education Governance & Policy*, 1: 18–31.

Sousa, C. A. A., and Hendriks, P. H. J. (2007) 'That Obscure Object of Desire: The Management of Academic Knowledge', *Minerva*, 45: 259–74.

Stephan, P., Veugelers, R., and Wang, J. (2017) 'Blinkered by Bibliometrics', *Nature*, 544: 411–2.

van der Wall, E. E. (2012) 'Journal Impact Factor: Holy Grail?', *Netherlands Heart Journal*, 20: 385–6.

van Leeuwen, T. N., van der Wurff, L. J., and de Craen, A. J. M. (2007) 'Classification of 'Research Letters' in General Medical Journals and Its Consequences in Bibliometric Research Evaluation Processes', *Research Evaluation*, 16: 59–63.

Weingart, P. (2005) 'Impact of Bibliometrics upon the Science System: Inadvertent Consequences?', *Scientometrics*, 62: 117–31.

Whitley, R. 1984. *The Intellectual and Social Organization of the Sciences*. Oxford: Clarendon Press.

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

Wilsdon, J., Allen, L., Belifiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Johnson, B. (2015) *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE. DOI: 10.13140/RG.2.1.4929.1363, https://responsiblemetrics.org/the-metric-tide/.

Wouters, P. (1999) 'Beyond the Holy Grail: From Citation Theory to Indicator Theories', *Scientometrics*, 44: 561–80.