RESEARCH ARTICLE

Citation metrics covary with researchers' assessments of the quality of their works

Dag W. Aksnes, Fredrik Niclas Piro, and Lone Wanderås Fossum

Nordic Institute for Studies in Innovation, Research, and Education (NIFU), Oslo, Norway

Keywords: bibliometric indicators, citations, metrics, peer review, research quality, scientific importance

ABSTRACT

For a long time, citation counts have been used to measure scientific impact or quality. Do such measures align with researchers' assessments of the quality of their work? In this study, we address this issue by decomposing the research quality concept into constituent parts and analyzing their correspondence with citation measures. The focus is on individual publications, their citation counts and how the publications are rated by the authors themselves along quality dimensions. Overall, the study shows a statistically significant relationship for all dimensions analyzed: solidity, novelty/originality, scientific importance and societal impact. The highest correlation is found for scientific importance. However, it is not very strong, but we find distinct gradients when publications are grouped by quality scores. This means that the higher the researchers rate their work, the more they are cited. The results suggest that citation metrics have low reliability as indicators at the level of individual articles, but at aggregated levels, the validity is higher, at least according to how authors perceive quality.

1. INTRODUCTION

Citation data are widely used in the context of research evaluation and performance assessments (Wilsdon, Allen et al., 2015). How often a publication is cited in the research literature is seen as a sign of its valuation as a scientific contribution. The citation counts of individual publications constitute the basis for measuring performance at various levels in the research system, such as individual authors, research groups, departments, and institutions. The idea that citation numbers can be used as a proxy for research quality dates back a long time (Cole & Cole, 1971). Today, citations are still claimed to reflect research quality (Caon, Trapp, & Baldock, 2020), although most bibliometric professionals would probably adhere to the view that citations reflect scientific impact rather than quality.

A recent literature review examined the relationship between research quality and citation indicators (Aksnes, Langfeldt, & Wouters, 2019). The point of departure is the multidimensional character of the research quality concept, where plausibility/reliability, originality, scientific value, and societal value are seen as key characteristics. In Polanyi's original elaborations (1962), the merit of a scientific contribution relates to the three first dimensions, but societal value has been added by other scholars (Lamont, 2009; Weinberg, 1963). These key distinctions of research quality reappear in many later empirical studies (Langfeldt, Nedeva et al., 2020). Plausibility/reliability may refer to the solidity of empirical evidence, the soundness of the results, and their reliability; originality to providing new knowledge and innovative research; scientific value to the contribution to research progress and importance for other

an open access journal

Check for updates

Citation: Aksnes, D. W., Piro, F. N., & Fossum, L. W. (2023). Citation metrics covary with researchers' assessments of the quality of their works. *Quantitative Science Studies*, *4*(1), 105–126. https://doi.org/10.1162/qss_a_00241

DOI: https://doi.org/10.1162/qss_a_00241

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00241

Received: 15 September 2022
Accepted: 28 December 2022

Corresponding Author:
Dag W. Aksnes
dag.w.aksnes@nifu.no

Handling Editor:
Ludo Waltman

The MIT Press

research; and societal value to the usefulness for society. Although a multitude of other notions and aspects of research quality have been suggested, these can generally be regarded as specific cases of the four dimensions (Langfeldt et al., 2020).

The review by Aksnes et al. (2019) argues that citations, to some extent, indicate scientific impact and relevance, but there is little evidence of citations reflecting other key dimensions of research quality described above. The latter conclusion was based on an examination of the literature, showing that studies addressing the issue empirically are lacking.

The lack of previous studies addressing this issue is the motivation for the current paper. The aim is to provide further knowledge on the extent to which citations reflect the various dimensions of research quality. The focus is on individual publications, their citation counts, and how the publications are rated by the authors themselves along quality dimensions. Specifically, the following research questions are addressed:

- To what extent do citation metrics of publications correspond with the authors' self-assessments of research quality dimensions: novelty/originality, solidity, scientific importance, and societal impact?
- As a subordinate issue: To what extent does the relationship differ by type of research contribution (theoretical, empirical, methodological and reviews) and by research field?

The latter questions have been added because previous research has shown that citation patterns differ across types of contributions. Review articles are particularly known to be, on average, more frequently cited than ordinary articles (Mendoza, 2021; Miranda & Garcia-Carpintero, 2018). Moreover, some of the world's most highly cited publications are method papers (Aksnes et al., 2019; Small, 2018). Less is known about the citation scores of other types of contributions. However, a study by Aksnes (2006) showed relatively small differences in citations to theoretical, methodological, and empirical contributions. The field dimension is also important. Not only do citation patterns differ significantly across fields, but there are also large variations in the coverage of the scientific and scholarly literature in bibliometric databases (Aksnes & Sivertsen, 2019; Marx & Bornmann, 2015). This limitation particularly affects the humanities field as well as many social sciences disciplines, presumably affecting the validity of citation measures in performance analyses in these disciplines. Accordingly, it has been recommended that citation analyses be applied with caution in these areas (Moed, 2005; Ochsner, Hug, & Galleron, 2017). Therefore, our analysis will specifically address how correspondence differs across fields.

Considering the frequent use of citations and other publication-based metrics for research evaluation purposes, hiring, and funding (Langfeldt, Reymert, & Aksnes, 2021), the research questions of our study are important and should be paid attention to. Many studies have addressed similar research questions, comparing citation measures with external benchmarks (e.g., peer reviews). The results are typically interpreted within a validation framework, meaning that if citation indictors can legitimately be used as performance measures, there should be a certain congruity with peer assessments. For example, Harzing (2018) analyzed the British national research evaluation (Research Assessment Exercise [REF]) and compared universities' REF scores with the number of citations, reporting a very high correlation (0.97). However, the degree of correspondence identified differs significantly across individual studies and generally tends to be moderate and far from perfect (Aksnes et al., 2019; Wilsdon et al., 2015).

In an examination of the lack of consistency in previous REF-based comparative assessments, Traag and Waltman (2019) emphasized that the results will differ according to the level of aggregation studied, from individual publications to aggregated levels, such as institutions.

In this study, we focus on the lowest level of aggregation: individual publications. At this level, previous studies seem to have found rather low correspondence. One of the most comprehensive studies is the one carried out for the Metric Tide report (Wilsdon et al., 2015) on how REF 2014 quality scores correlated with metrics (Higher Education Funding Council for England (HEFCE), 2015). Here, the REF quality scores were based on peer assessments of the originality, significance, and rigor of the publications. A variety of indicators were examined, but none obtained a higher correlation coefficient (Spearman's) than 0.34. Similarly, a recent study of bibliometric indicators and peer reviews of the Italian research assessment showed weak correspondence at the level of individual articles (Baccini, Barabesi, & De Nicolao, 2020), concluding that metrics should not replace peer reviews at the level of individual articles. This was based on a combined index in which the number of citations and journal impact factors were used. Other examples include a study analyzing articles that were singled out in Mathematical Reviews as being especially important (Smolinsky, Sage et al., 2021). Of these, 17% were highly cited (among the top 1% cited papers). Articles that have been recommended as important in biomedicine by the Faculty of 1,000 (a publication peer-review service for biological and medical research) have also been shown to correlate with citations, but only weakly (Waltman & Costas, 2014). Borchardt, Moran et al. (2018) analyzed chemistry articles and found that peer assessments of importance and significance differed considerably from citation-based measurements. Older studies of individual articles with similar findings include Aksnes (2006) and Patterson and Harris (2009), but there are also studies that have concluded differently. Ioannidis, Boyack et al. (2014) found that papers ranked by elite scientists as their best were also among the most highly cited. Similarly, an examination of award-winning papers in economics showed that these papers had a significantly higher number of citations than ordinary papers (Coupe, 2013).

The large variety in the observed degree of correspondence in previous comparative studies may not be surprising. Not only is research quality a multidimensional concept, but peer evaluations also often include assessments of factors besides quality. Thus, the foundation for simple comparative assessments may be weak or lacking. Moreover, many citation indicators exist, and the results may depend on the type of indicator selected. Finally, peer assessments are uncertain and fallible (Aksnes, 2006; Traag & Waltman, 2019).

Against this background, we believe there is a need for more studies that address the topic in a simpler and more transparent manner. In our view, a problem or limitation with many previous studies is that the multidimensional character of research quality is not taken into account. This paper expands the perspective by decomposing the concept, making it evident which research performance or quality dimensions are compared with citation metrics.

In the study, we rely on authors' self-assessments of their papers' quality dimensions, which is an approach also adopted in several previous studies (Aksnes, 2006; Case & Higgins, 2000; Dirk, 1999; Ioannidis et al., 2014; Porter, Chubin, & Xiao-Yin, 1988; Shibayama & Wang, 2020). Still, there are pros and cons to such a methodology. A main advantage is that the authors have thorough knowledge of the content of the publication, the research reported, and the field. However, their views may be regarded as more subjective than those of their peers. For example, one might expect certain psychological mechanisms at play, such as the Dunning–Kruger effect (Kruger & Dunning, 1999), which states that people overestimate their own abilities but where the effect is reversed for highly skilled individuals. At the same time, there are also limitations to alternative approaches that rely on peer assessments. As noted above, these are fallible, and the agreement between different reviewers has been shown to be very low (Lee, Sugimoto et al., 2013), meaning that there is no objective yardstick to which citations can be validated as indicators.

2. DATA AND METHODS

2.1. Study Design and Questionnaire

Citation distributions are skewed at the level of individual authors (Seglen, 1992). A large number of published articles are little cited or not at all. Rather than selecting a random set of publications and authors, which would be dominated by less-cited publications, we designed a method for which contributions from a wide citation spectrum would be well represented. Specifically, this means that we divided the publications into three categories:

- Highly cited publications (within the top 10 percentile rank)
- Publications with intermediate number of citations (within the top 10–50 percentile rank)
- Less-cited publications (within the bottom 50–100 percentile rank)

We then preselected individuals who had published at least one article in each of the groups during the period analyzed and where they appeared as either the first or last author. The latter criterion was added because we would like to include publications in which the authors had contributed in key roles. Usually, this is indicated by first authorship (main contributor) or last authorship (principal investigator), although this rule does not always hold, as alphabetical author listing is common in some fields (Waltman, 2012).

Furthermore, the study was limited to scientists affiliated with institutions in Norway. One might ask whether this specific national delimitation has relevance when interpreting the results. On the one hand, the attributes of research quality analyzed are thought to be universal in the way that they transcend specific field or national delimitations. On the other hand, these attributes might still be given different content in different contexts (Langfeldt et al., 2020). Researchers' perceptions of quality may also be influenced by national research evaluation systems (Bianco, Gras, & Sutz, 2016). Still, we do not think national peculiarities should be given much emphasis when interpreting the results. A large majority of the publications also have coauthors from other countries and do not therefore represent "domestic" Norwegian research.

What should be emphasized is that the investigation is not based on a random selection of individuals. The survey is biased in favor of scientists who have published highly cited papers, have key roles in research, and are reasonably productive. In practice, this means that the survey is dominated by experienced scientists, often in full-professor positions.

A questionnaire was designed in which the authors were asked questions about three of their publications, one randomly selected from each of the citation categories described above. The respondents were not informed about this strategy, as their responses should not be influenced by knowing the citation-based selection procedure. We therefore simply asked them to rate three of their papers that had been randomly selected.

The questions were identical for all papers. Specific questions were included for each of the different quality dimensions, in addition to a general question on the type of contribution. We also included a question on groundbreaking research, as this has been claimed to be associated with highly cited publications (Savov, Jatowt, & Nielek, 2020). An overview is provided in Table 1.

As can be seen from Table 1, the various research quality dimensions were operationalized only to some extent in the survey, leaving some room for the respondents' own interpretations. The operationalization is based on an examination of the relevant literature on research

Table 1. Overview of questions and answer alternatives

Questions	Answer alternatives
Please characterize the main contribution(s) from these articles	Theoretical–Empirical – Methodological–Review–Cannot say/Not relevant
How do you regard the novelty/originality of the research reported in these articles (e.g., of in terms of topic addressed, research question, methodology and results)?	1 (low)–2–3–4–5 (high)-Cannot say/Not relevant
How do you regard the solidity of the research reported in these articles (i.e., validity and certainty/reliability of the methods and results reported)?	
How do you regard the scientific importance of the research reported in these articles (e.g., in terms of new discoveries/findings, theoretical developments and new analytic techniques)?	
As far as you know, has the research/results presented in these articles had any societal impact (i.e., effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment, or quality of life, beyond academia)?	Yes–No–Don't know
Would you consider your article as 'groundbreaking research'?	

quality and research evaluations; specifically, the one on societal impact relies on the definition applied in REF (Savov et al., 2020).

2.2. Bibliometric Data, Indicators and Analyses

The study relies on two bibliometric databases. The first is the Norwegian publication database, Cristin, which contains complete data on the scientific and scholarly publication output of Norwegian researchers (Sivertsen, 2018). From this database, the publication outputs of individual researchers can easily be identified. The second is the Web of Science (WoS) database, which has been used to retrieve citation data. The two databases were coupled through a joint identity key. We applied a local version of WoS maintained by the Norwegian Agency for Shared Services in Education and Research. Thus, the study is limited to articles that have been indexed in WoS.

We identified the publication output of all Norwegian researchers (covering higher education institutions, hospitals, and independent research institutes) for the period 2015–2018. We did not include more recent publications, as we required a citation window of at least three years. Moreover, we did not include older publications, as the memory of the respondents may be more limited when going back in time.

Only publication items classified as regular articles in WoS were included. We excluded review articles (because the survey focused on the characteristics of original or primary research) as well as minor items, such as editorials and letters. Nevertheless, we preserved the review category in the questionnaire because the WoS item classification system is known to be inaccurate (Donner, 2017).

Two types of citation indicators were used. First, we used the normalized citation index (MNCS), where the citation numbers of each publication are normalized by subject field, article type and year (Waltman & Van Eck, 2013), thus allowing publications from different fields and years to be compared on equal grounds. Second, we used the citation percentile, ranging

from 0 to 100%. This is an indicator showing the articles' position within the citation distribution of their field (Waltman & Schreiber, 2013), also taking into account their publication year and article type. Thereby, we cover the two most commonly applied indicators in citation studies in which the percentile-based indicator seems to be increasingly preferred due to its mathematical properties and insensitivity to outliers (Hicks, Wouters et al., 2015; Wilsdon et al., 2015).

The survey results are combined with data on the percentile category of the publications through descriptive bivariate analyses. In this way, we used a binning approach. Although this implies the loss of some information, it simplifies the analyses and the visual interpretation of the results. The most common approach in previous similar studies is correlation analysis (see, for example, Aksnes, 2006; Borchardt et al., 2018; Smolinsky et al., 2021). In this study, we carried out supplementary analyses using Pearson's correlation coefficient in the case of the percentile citation indicator and Spearman's rank correlation coefficient in the case of the MNCS. Spearman's test was used in the latter case due to the lack of normally distributed data. Thus, the strength of the relationship between a ranking derived from the MNCS indicator and author ratings is analyzed.

2.3. The Survey

The survey was distributed in January 2022 using SurveyXact software. Questionnaires were sent out to a sample of 1,250 researchers based on the criteria described above. The response rate was 47%, and the final sample consisted of 592 individuals, each with three publications included. The study, therefore, encompasses assessments of almost 1,800 publications.

Table 2 shows how the respondents are distributed by scientific domain and gender. In total, 180 women (30.4%) and 412 men (69.6%) participated in the survey. Medicine and Health is the largest domain (43.4%), and there were few participants from the Humanities (4.6%). This is due to the publication patterns of the Humanities, where only a small part of the publication output is indexed in WoS (Aksnes & Sivertsen, 2019). We acknowledge that the number of respondents from Humanities is low. The results in this domain should accordingly be treated with caution (especially as many researchers do not publish in WoS-indexed journals, but rather in national language journals and in books). The field classification applied in the study relies on the system of the Cristin database (Sivertsen, 2018), where each publication is assigned field categories and several researchers have publications in two (and even three) domains. In Table 2, they are listed according to the majority principle.

Data on the birth years of the researchers (not shown in tables) show that the average ages of the male and female respondents are 56.8 years and 54.7 years, respectively. Thus, we are dealing with an experienced group of researchers, which is also evident from the figures on the

Table 2. Distribution of respondents by scientific domain and gender

Scientific field	Women	Men	Total	% Women	% Men	% Total
Humanities	10	17	27	5.6	4.1	4.6
Medicine and Health	92	165	257	51.1	40.0	43.4
Natural Sciences and Technology	39	159	198	21.7	38.6	33.4
Social Sciences	39	71	110	21.7	17.2	18.6
Total	180	412	592	100	100	100

total publication output of the respondents during the period, which is 13 publications per year on average. Here, women have a substantially lower number than men: 10 versus 14 publications. The skewed distribution of men and women across fields (both vertically and horizontally in Table 2) is quite similar to that seen in the total Norwegian population of researchers in senior positions (e.g., Nygaard, Aksnes, & Piro, 2022).

To assess possible response biases, we compared the field, gender, and age distributions of the respondents with those for the original sample. The response rate was 44% for women and 49% for men. The respondents were slightly older than the nonrespondents (average age +3 years). At the level of domain, the response rate was lowest in the two largest fields (Medicine and Health, 44%, and Natural Sciences and Technology, 46%), somewhat higher in Social Sciences (55%) and much higher in Humanities (71%). Hence, the differences in response rates across domains have led to a more balanced representation by reducing some of the original size differences. We consider the response bias across the two other variables as minor, not representing a methodological problem.

3. RESULTS

3.1. Type of Contribution

Table 3 shows the distribution of the publication types (self-reported) across three citation rank categories: top 10 percentile (highly cited publications), 10–50 percentile and the 50–100 percentile (less-cited/uncited publications). For almost half of the publications (47%), the main contribution was assessed to be empirical. One-quarter of the papers were assessed as foremost contributing theoretically, and 20% contributed methodologically. As noted above, articles classified as reviews in WoS were excluded from the sample. Nevertheless, 7% of the papers were claimed by the authors to foremost have "review" as the main contribution. At the level of domain, theoretical contributions appear more frequently in the Social Sciences and Humanities, while empirical contributions dominate in the Medicine and Health sciences.

The distribution of articles across citation rank categories did not differ much (Table 3). Thus, the results do not suggest that certain types of contributions tend to be more highly cited than others. Even the review contributions are distributed quite evenly across citation categories. This might seem surprising, as review papers have generally been shown to be more highly cited (Miranda & Garcia-Carpintero, 2018). However, the data set was preselected not to include review articles, thus preventing any conclusions on this matter.

Table 3. Distribution of publications by type of contribution and citation rank categories (%)*

Type of contribution	0–10%	10–50%	50–100%	Total
Theoretical	26.2	25.8	25.5	25.9
Empirical	46.0	48.3	47.3	47.2
Methodological	19.9	20.0	20.7	20.2
Review	7.9	5.9	6.4	6.8
Total	100	100	100	100
N	592	592	592	1,776

* The respondents were allowed to select more than one type of contribution; therefore, double counts occur. Missing values and 'don't know' replies are excluded from the calculations.

Overall, the results imply that the contribution type variable is of little interest to include in forthcoming analyses, which would otherwise have been the case if a contribution type was over- or underrepresented in particular citation rank categories.

Below, we turn to the various components of the research quality concept.

3.2. Solidity

The respondents were asked to assess the solidity of the research reported in their publications from 1 to 5. The lowest score options (1–2) were used to a very small extent; this also holds true for the other survey questions. Overall, the respondents assessed solidity as moderate or high. The distribution by citation rank categories is shown in Figure 1.

Across all citation intervals, the large majority of the publications were assessed to have high solidity (scores 4 and 5). Even the less-cited publications were usually considered to have high solidity. However, the distribution of solidity assessments was not equal across the citation groups. A larger proportion of the most cited papers (10%) obtained the highest score (5) (54%), compared with 40% for the papers in the 50–100 percentile category. We also observe the opposite pattern for papers with the lowest score (1–3), amounting to 7% and 21%, respectively. In sum, Figure 1 gives the impression that perceptions of solidity increase with citation scores.

To analyze the correspondence at field levels, we applied a simplified approach calculating the average author score across citation rank categories (Figure 2). Except for the Humanities, we observe that the scores increased according to citation rank categories.

3.3. Novelty/Originality

On the question concerning the novelty/originality of the research, only 5.4% of the articles were rated with the lowest scores (1 and 2) (Figure 3). Approximately 25% obtained an intermediate score of 3, while the remaining articles were classified in the two highest categories (4–5). The pattern is quite similar to that observed for solidity, but the respondents were somewhat more modest when assessing novelty/originality. For example, although 46.3% of the articles were rated 5 on solidity, the corresponding figure for novelty/originality is 29.8%.

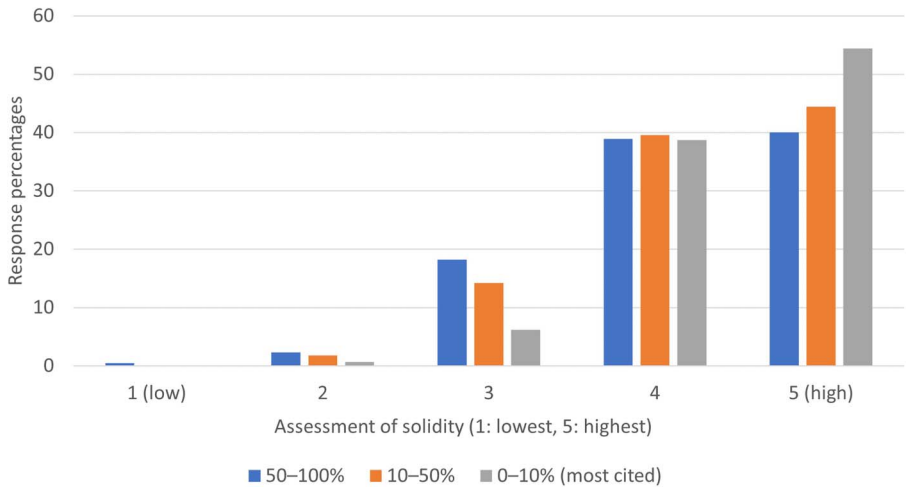

Figure 1. Distribution of publications by solidity score and citation rank categories (%).

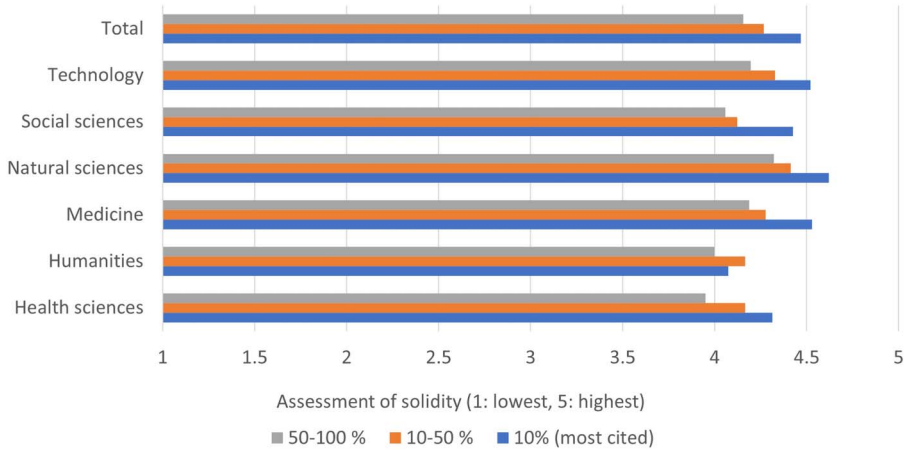

Figure 2. Average solidity score by citation percentile categories and scientific domain.

The distribution by citation rank categories shows that in the top 10 percentile, 43% of the articles got the highest score. The corresponding figure for the articles in the 50–100 percentile is 19%. Similarly, we see an opposite pattern for articles in the 50–100 percentile category, where more of the publications got low/intermediate scores (1–3). Thus, we see a tendency for the ratings to correspond with the citation rank categories of the publications.

Figure 4 shows the mean and total scores on novelty/originality across fields. In all fields, articles in the top 10 percentile are ranked highest. Similarly, articles in the 50–100 percentile category clearly have lower scores than those in the 10–50 percentile category, with the exception of Medicine.

3.4. Scientific Importance

The third dimension related to scientific quality is scientific importance. The largest number of articles were rated 4 (39%), with equal shares rated 3 or 5 (27%) (Figure 5). The distribution is very similar to the one previously shown for novelty/originality. The interpretation of this is that the researchers acknowledge that although the work itself is of high solidity, it may not have been equally novel/original or scientifically important.

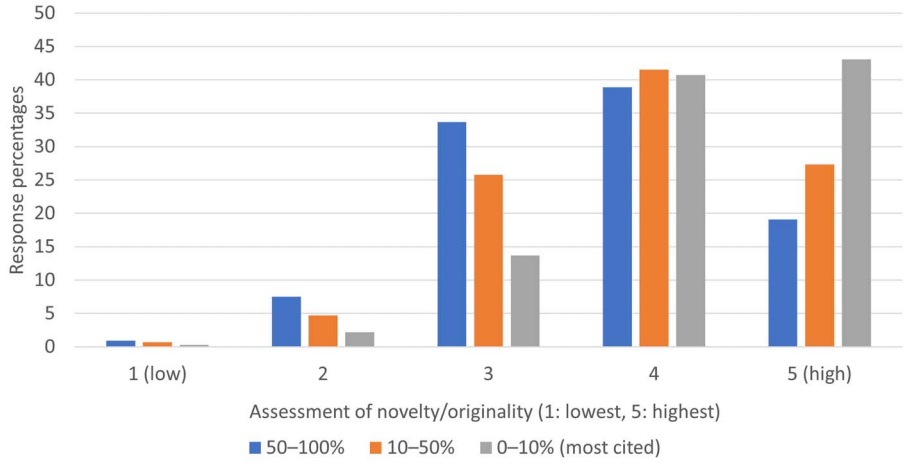

Figure 3. Distribution of publications by novelty/originality score and citation rank categories (%).

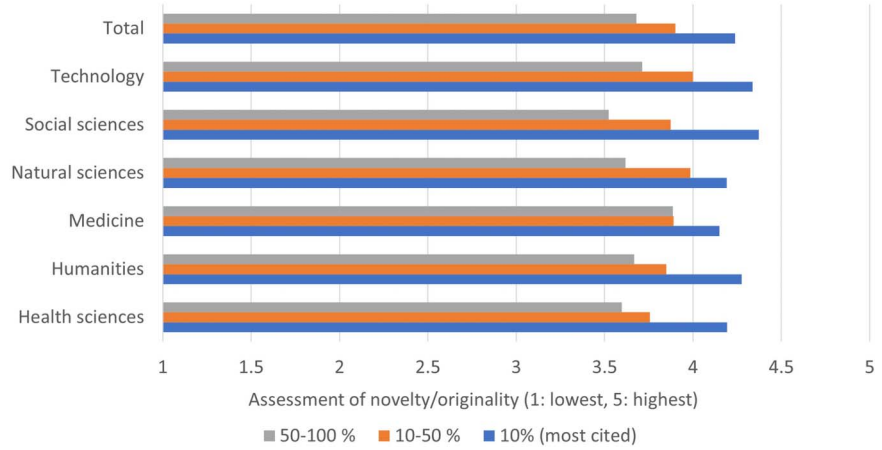

Figure 4. Average novelty/originality score by citation percentile categories and scientific domain.

The distribution by citation rank categories is also very similar to the one we observed for novelty/originality. A much higher proportion of the top 10 percentile articles obtained the highest score (45%) compared with articles in the 50–100 percentile category (14%).

Figure 6 shows the mean and total scores for scientific importance across fields. We do not observe large differences here. In all fields, the patterns are quite similar.

In sum, the differences in scores for the three quality dimensions indicate that the researchers have been able to differentiate between the dimensions (i.e., they have not automatically scored each paper with equal rating for all dimensions). This was confirmed with, first, a correlation analysis, revealing correlations (Pearson's *r*, two-tailed, sig. 000) of .350 between novelty/originality and solidity; .651 between novelty/originality and scientific importance; and .408 between solidity and scientific importance. Second, we calculated the percentages of scores that differed between pairs of quality dimensions. In 54.5% of the papers the researchers rated novelty/originality and solidity differently. For novelty/originality and scientific importance different scores were given for 41.5% of the papers, and for solidity and scientific importance, different scores were given for 56.9% of the papers.

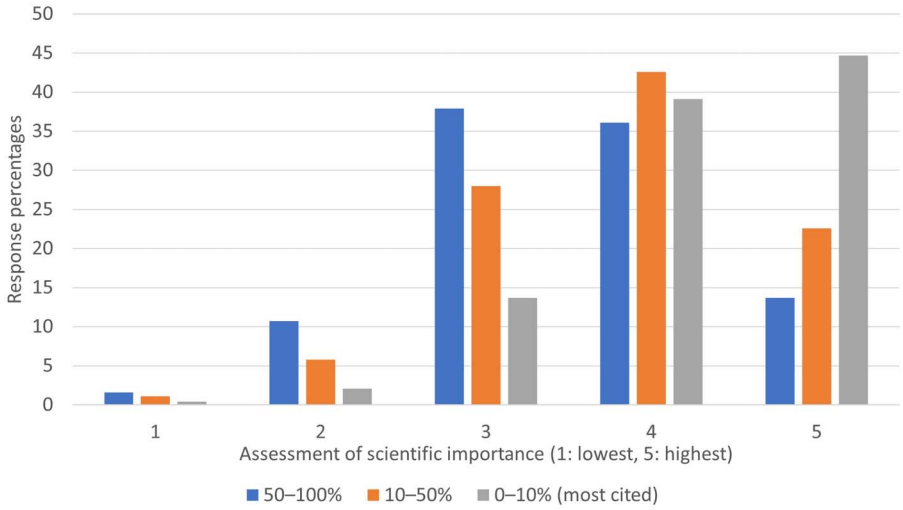

Figure 5. Distribution of publications by scientific importance score and citation rank categories (%).

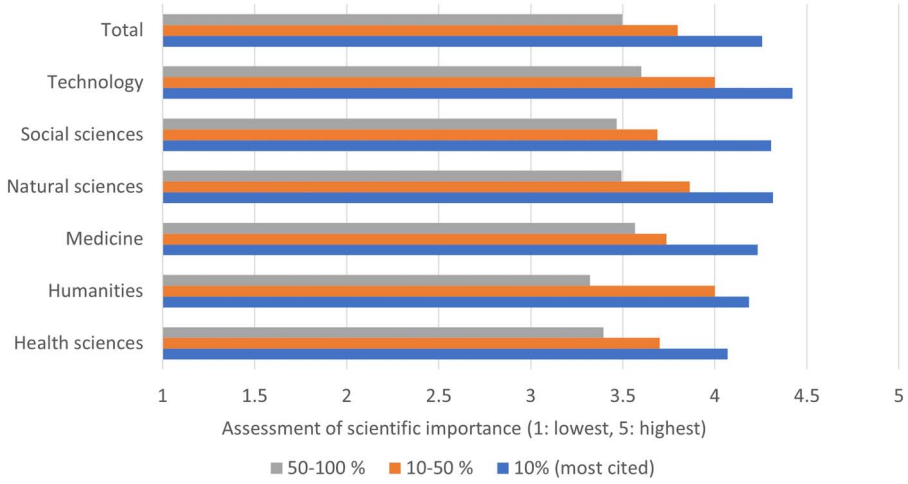

Figure 6. Average scientific importance score by citation percentile categories and scientific domain.

3.5. Groundbreaking Research

The researchers were also asked whether their publications represented "groundbreaking research." In total, more than one-quarter of the papers were perceived as groundbreaking (Figure 7). In the Humanities, this proportion was as high as 39.5%. In the Health Sciences the researchers reported that their papers were groundbreaking substantially less often (16.3%). In the other domains, the percentage ranged from 23.8% (Medicine) to 31.4% (Technology).

The respondents' assessments corresponded well with the citation range categories. A much higher proportion of the 10 percentile articles were considered groundbreaking compared with the two other categories. Still, 15% of the articles in the lowest 50–100 percentile category were considered groundbreaking. However, not all highly cited publications were considered groundbreaking. In all fields except Medicine, there was a distinct pattern corresponding with increasing proportions from 50–100, to 10–50, to the 10 percentile, but in Medicine, the reporting of groundbreaking research is twice as high in the 10 percentile compared to the other percentiles.

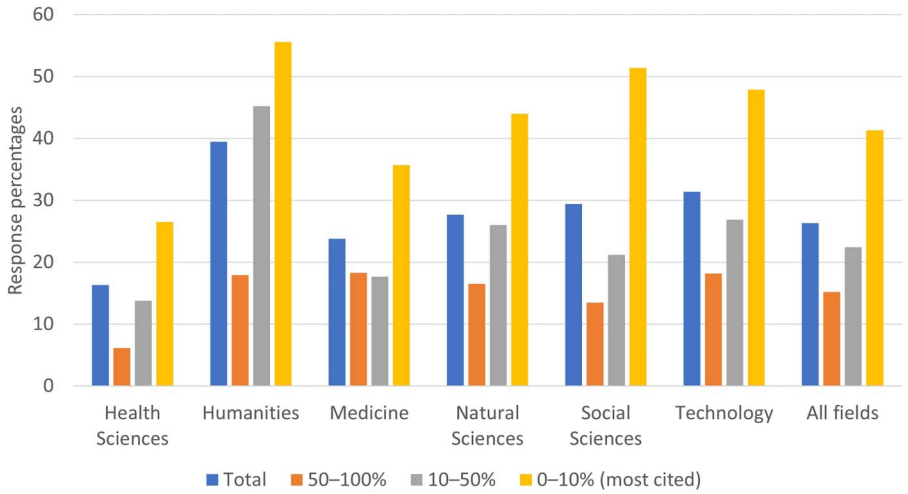

Figure 7. Percentage of publications considered to be groundbreaking research by citation percentile categories and scientific domain.

At the level of individuals (excluding those who did not answer the question for all three of their papers), 22 respondents (3.8%) claimed that all three of their papers were groundbreaking, while 256 respondents (44.7%) claimed that none of their papers were groundbreaking. Ninety-one respondents (15.9%) claimed two of their papers were groundbreaking and one was not. One groundbreaking paper and two nongroundbreaking papers were reported by 204 respondents (35.6%).

3.6. Societal Impact

Societal impact is the last dimension of the research quality concept. Here, the response alternatives were simply "yes," "no," and "don't know." A total of 24.9% of the papers were claimed to have had societal impact (Figure 8). In this calculation, the "don't know" publications are also included in the denominator because it is fully possible to know that your research has had societal impact, but it is not possible to know with certainty that it has not (the researcher may simply not be aware of it), which makes it not so important to distinguish between "no" and "don't know."

There are notable differences across fields in the extent to which the respondents consider their research to have societal impact. This proportion is highest in Medicine (31%) and the Social Sciences (29.8%) and lowest in Humanities (11.5%) and Health Sciences (12.7%).

In addition, in this dimension, the respondents' answers correspond with the citation rank categories in the previously observed manner: The highest proportion is for the 10 percentile group (33.9%) and lowest for the 50–100 percentile group (17.1%). This pattern is consistent across all fields.

3.7. Further Analyses

We have so far presented results using a binning approach consisting of three citation percentile categories corresponding to the criteria applied in the selection of publications. However, some information is lost by this procedure, and we will present analyses using accurate data (citation scores). In addition to the percentile-based indicator, we will analyze the MNCS.

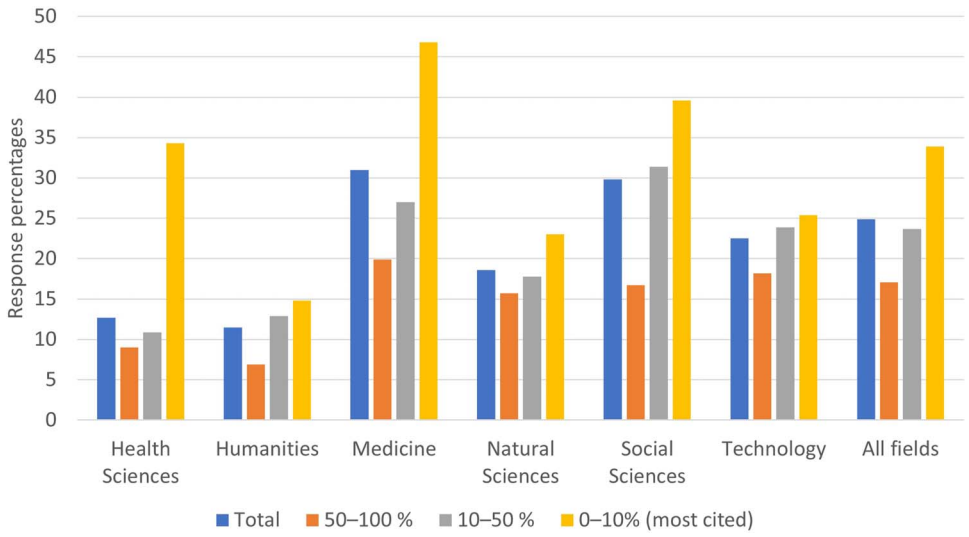

Figure 8. Percentage of publications considered to have had a societal impact across citation rank categories and scientific domains.

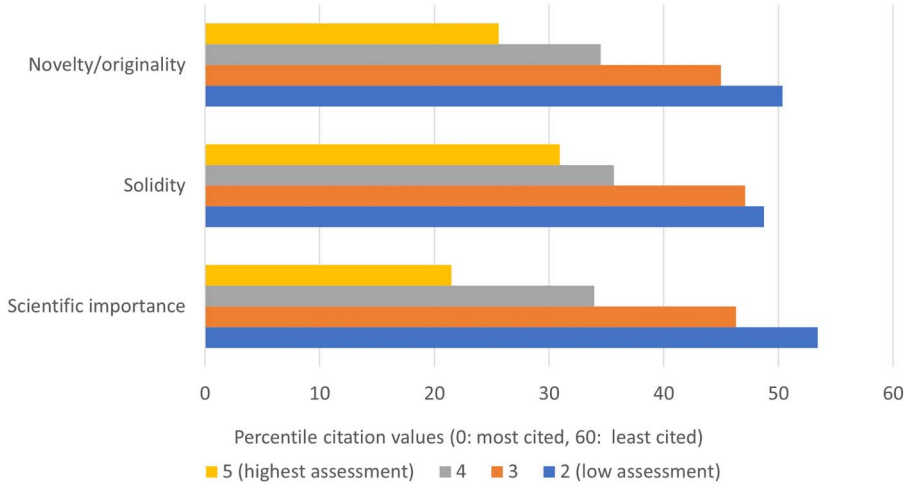

Figure 9. Percentile citation values and assessments of research quality dimensions. Response alternative 1 (lowest) is not shown due to a very small number of observations.

Figure 9 shows how the respondents' assessments of quality dimensions vary according to response alternatives and percentile values. For example, the average percentile value for articles rated with high (5) novelty was 26, compared with 50 for articles rated with the lowest score (2). For all quality dimensions, the results correspond with the patterns identified above, in which there is a distinct difference in values by response alternatives.

We then carried out a similar analysis using the other citation indicator, MNCS (Figure 10). A corresponding pattern is found but with a larger difference in citation values across the scores (1–5). This is due to the distributional character of the MNCS indicator and the presence of outliers.

Further insights concerning the relationships are obtained by carrying out correlation analyses. Both MNCS and percentile values show moderate to weak correlations with the three quality dimensions. However, the associations are statistically significant, as shown in

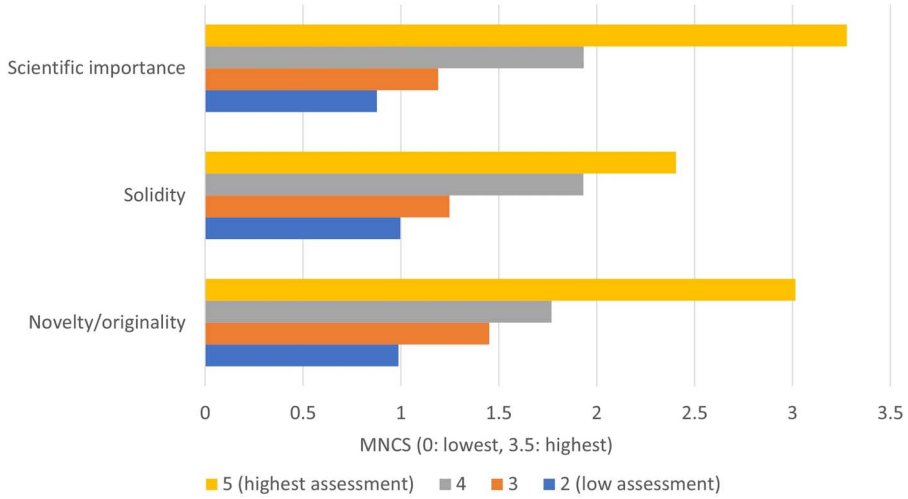

Figure 10. Mean normalized citation score (MNCS) and assessments of research quality dimensions. Response alternative 1 (lowest) is not shown due to a very small number of observations.

Table 4. Correlation analysis of MNCS/citation percentiles and research quality measures

		Solidity	Novelty/originality	Scientific importance
MNCS	Spearman's correlation	.177*	.280*	.370*
	Sig. (2-tailed)	.000	.000	.000
Percentile values	Pearson correlation	−.181*	−.266*	−.351*
	Sig. (2-tailed)	.000	.000	.000
	N	1,691	1,732	1,707

* Correlation is significant at the 0.01 level (2-tailed).

The different signs reflect that a low percentile value corresponds with high citation counts and vice versa.

Table 4. There are minor differences only across the two types of citation indicators. The strongest correlation is between citation measures and scientific importance.

A more differentiated picture emerges when we analyze correlations separately by field. Table 5 shows the results for the citation percentile indicator. Very similar patterns were

Table 5. Correlation analysis (Pearson's *r*) of citation percentiles and research quality measures across scientific fields

		Solidity	Novelty/originality	Scientific importance
Health sciences	Correlation	−.194**	−.318**	−.338**
	Sig. (2-tailed)	.002	.000	.000
	N	257	268	262
Humanities	Correlation	−.085	−.219*	−.341**
	Sig. (2-tailed)	.438	.036	.001
	N	85	92	86
Medicine	Correlation	−.203**	−.111*	−.290**
	Sig. (2-tailed)	.000	.020	.000
	N	428	435	428
Natural sciences	Correlation	−.182**	−.278**	−.378**
	Sig. (2-tailed)	.000	.000	.000
	N	424	426	426
Social Sciences	Correlation	−.215**	−.418**	−.407**
	Sig. (2-tailed)	.000	.000	.000
	N	293	307	302
Technology	Correlation	−.173*	−.311**	−.379**
	Sig. (2-tailed)	.013	.000	.000
	N	204	204	203

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

obtained for the MNCS indicator (not shown). Contrary to our expectations, there were rather small differences across fields. The rather weak correlation regarding solidity (Table 4) is also observed for most fields, but the variety between fields on the other quality dimensions shows that correlations range from very small on novelty/originality (Medicine: −.11) to moderate (Health Sciences: −.32; Social Sciences: −.42). For scientific importance, all correlations were in the range of −0.29 to −.41.

In Table 6, we show the mean and median values of MNCS for each research quality dimension and score. The median values are clearly lower than the average, demonstrating a skewed distribution in which the mean values are strongly influenced by a relatively small group of very highly cited papers. In almost all fields, there is a distinct gradient, meaning that the citation scores increase according to the self-assessments of the papers. For scientific importance, the gradient is congruent with author ratings in all fields except Technology, where papers with a score of 3 have lower MNCS than those with a score of 2. For the other dimensions, there are also very few exceptions to the overall pattern, in which citation scores increase with the respondents' scores.

To test for the impact of the skewed data distribution in responses to the quality dimensions (with a dominance of respondents answering at the higher end of the scale), we ran

Table 6. Mean and median values of MNCS across fields and research quality dimensions*

		2 (low)		3		4		5 (highest)	
Indicator	**Field**	**Avg.**	**Median**	**Avg.**	**Median**	**Avg.**	**Median**	**Avg.**	**Median**
Novelty/originality	Health sciences	0.66	0.48	1.24	0.73	2.17	1.75	4.91	2.63
	Humanities	–	–	2.30	1.42	1.91	1.13	4.12	3.22
	Medicine	1.45	1.10	1.75	0.84	1.65	1.27	2.44	1.16
	Natural sciences	1.11	0.52	1.38	0.60	1.40	0.82	2.74	1.99
	Social sciences	0.68	0.41	1.21	0.46	1.99	1.34	3.26	2.94
	Technology	0.94	0.60	1.24	0.57	1.62	1.08	2.21	2.27
Solidity	Health sciences	0.35	0.44	1.45	0.77	2.00	1.41	3.86	1.58
	Humanities	–	–	2.69	1.48	2.80	0.99	2.31	1.62
	Medicine	1.48	0.86	1.07	0.70	1.95	0.89	2.35	1.56
	Natural sciences	1.08	1.13	0.74	0.48	1.80	0.89	2.00	1.28
	Social sciences	–	–	1.09	0.50	1.94	1.35	2.68	1.83
	Technology	–	–	1.41	0.72	1.58	0.92	1.89	1.58
Scientific importance	Health sciences	0.97	0.44	1.41	0.84	2.39	1.63	5.08	2.62
	Humanities	1.38	0.33	1.49	0.35	2.96	1.73	3.50	2.59
	Medicine	0.73	0.51	1.28	0.84	1.83	0.95	3.00	2.26
	Natural sciences	0.97	0.39	0.96	0.55	1.64	1.06	3.06	2.49
	Social sciences	0.58	0.39	1.23	0.48	1.96	1.20	3.55	2.97
	Technology	1.10	0.74	0.87	0.42	1.64	1.08	2.42	2.31

* Cells with fewer than 5 papers are not shown in the table. Cannot say/not relevant responses and missing values are also excluded.

Table 7. Independent-samples Kruskal–Wallis* test of MNCS and research quality dimensions

Samples compared	Quality dimension	Test statistic	Std. error	Std. test statistic	Sig.
2–1	Novelty	46.8	160.5	.33	.770
2–1	Solidity	–274.3	297.2	–.92	.356
2–1	Scientific importance	–24.2	128.8	–.22	.851
2–3	Novelty	–88.3	60.0	–1.51	.141
2–3	Solidity	–293.3	283.8	–1.03	.301
2–3	Scientific importance	–126.4	121.8	–1.04	.299
2–4	Novelty	–259.0	58.1	–4.54	.000
2–4	Solidity	–476.0	282.5	–1.77	.092
2–4	Scientific importance	–340.6	121.0	–2.81	.005
2–5	Novelty	–433.0	59.1	–7.32	.000
2–5	Solidity	–565.0	282.4	–2.00	.045
2–5	Scientific importance	–568.2	121.7	–4.77	.000
1–3	Novelty	–41.5	152.7	–.33	.786
1–3	Solidity	–19.0	99.6	–.22	.849
1–3	Scientific importance	–102.2	53.2	–1.92	.055
1–4	Novelty	–212.2	152.0	–1.43	.163
1–4	Solidity	–201.7	95.9	–2.10	.035
1–4	Scientific importance	–316.4	51.5	–6.14	.000
1–5	Novelty	–386.2	152.4	–2.54	.011
1–5	Solidity	–290.6	95.6	–3.04	.002
1–5	Scientific importance	–544.0	53.1	–10.20	.000
3–4	Novelty	–170.7	30.8	–5.54	.000
3–4	Solidity	–182.7	38.2	–4.88	.000
3–4	Scientific importance	–214.2	30.0	–7.15	.000

Table 7. (*continued*)

Samples compared	Quality dimension	Test statistic	Std. error	Std. test statistic	Sig.
3–5	Novelty	–344.7	32.8	–10.50	.000
3–5	Solidity	–271.6	37.5	–7.33	.000
3–5	Scientific importance	–441.9	32.6	–13.60	.000
4–5	Novelty	–174.1	29.0	–6.00	.000
4–5	Solidity	–88.9	25.8	–3.44	.001
4–5	Scientific importance	–227.7	29.8	–7.63	.000

* Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

nonparametric tests to determine statistically significant differences in MNCS between the scores on the quality dimensions. In all three dimensions, the Kruskal–Wallis *H* test rejected the null hypothesis (that the distribution of MNCS is the same across categories of quality dimensions, sig. .000), with MNCS values significantly different (and higher) between comparisons of groups scoring 3 and higher (Table 7) but with some nonsignificant results in comparison, including scores 1–2, where the number of responses is very low.

4. DISCUSSION

The main purpose of this study was to assess how citation indicators align with researchers' subjective assessments of research quality dimensions. In addition, the study has provided knowledge on how researchers evaluate their own research. We discuss some of the findings related to this.

Generally, researchers rate their research publications highly. More than one-quarter of the papers are perceived by the authors as groundbreaking research. Almost none of the publications are rated with the lowest quality score (1), few are rated with the second lowest score (2), and a large majority are rated with the two highest scores (4 and 5). This holds true across all quality dimensions assessed in the survey.

The pattern is particularly evident for solidity, while the researchers are slightly more restrained when assessing the novelty/originality and scientific importance of their research. This is perhaps not surprising, as studies that lack solidity would not be accepted for publication in reputable journals. In addition, studies with little novelty might be filtered out through the peer review process. Thus, by being accepted for publication in journals, the studies have already undergone a selection process that might explain the patterns.

The respondents of our survey are all experienced researchers, typically having a long career within the academic system; they have also contributed to highly cited publications. It is reasonable to assume that they would be reluctant to contribute to studies they consider having little novelty or scientific importance or lacking solidity. This is an additional factor that might explain why so few of the articles are ranked with low scores.

At the same time, researchers may be too positive about their research. The study shows how scientists think about their research, and in this way, there is prima facie reliability. However, an open question is to what extent their assessments would be shared by external reviewers.

Despite these two limitations, the judgments of the researchers provide an interesting and suitable reference point for addressing the validity of citations as performance measures. First, the spread of responses is still large enough to allow comparative analyses. Second, using author views is one approach in the range of studies applying different reference points, all with various strengths and limitations.

4.1. Possible Biases in the Respondents' Assessments

Citation metrics are available through a large number of products and services. Many researchers know their own metrics and which publications are highly cited. In Norway, citation analyses have also been provided in broad peer-based national evaluations of research fields. To what extent the respondents' answers are affected by beforehand knowledge, we simply do not know. We find it important to note that while some researchers' answers may be consciously or unconsciously guided by their knowledge of citation numbers, for others they may not, and some may even "disagree" with the citation scores (cf. Aksnes & Rip, 2009). One might also think that the perceived prestige of the publication channels has some influence, such as that a *Nature* paper automatically receives a high rating. Still, as we do not find strong correlations between citation metrics and author ratings, we do not think this is a major problem, and *prima facie* we do not see any reason why the respondents would rate their papers systematically in line with known citation numbers, rather than providing a sincere assessment.

Moreover, the results may be biased due to a lack of memory. Some researchers responded that the selected papers were not among their most relevant ones and that they contributed in peripheral roles only (making it difficult to "remember"). In particular, people who have been involved in a large number of papers in recent years may have limited memories of individual contributions. Additionally, one might ask if people will rate their more recent papers higher. However, when testing this issue, we did not find any difference at all on the various quality dimensions (i.e., the average scores for papers were identical across all 4 years).

4.2. Correlations Are Not Strong, But Significant for All Quality Dimensions

Overall, this study has disclosed notable covariations between citation indicators and author ratings of quality dimensions. Highly cited publications tend to obtain substantially higher ratings than less-cited publications and vice versa. This holds for all quality dimensions but is weaker for solidity than for the other. The latter finding reflects that most publications (85%) were assessed to have high solidity (scores 4 and 5) in the first place.

The agreement between citation indicators and author assessments has been analyzed using different approaches. In the first approach, based on binning, we looked at three main citation categories. Here, we observed a distinct pattern in which highly cited publications are generally seen as having a higher research quality than other publications. This holds true across the different dimensions analyzed. To the contrary, publications that have been moderately or little cited are ranked lower but rarely as of inferior quality. In this way, quality rankings of publications based on rough citation categories seem to have certain justifications—at aggregated levels.

In the other part of the analysis, exact citation figures were used as the basis. These analyses support the main findings of the first part. Articles that have been ranked as high quality are, on average, more cited. Comparing the averages in this way reveals distinct differences between the groups.

The results provide some support for the hypothesis derived from Aksnes et al. (2019) that citation rates reflect scientific impact and relevance (to some extent) but not the other quality dimensions. There is a significant correlation for all the quality dimensions analyzed. The correlation is strongest for scientific importance, which holds true for both the citation indicators analyzed: the percentile (−0.35) and the MNCS indicator (0.37). In particular, highly cited papers are considered to have high scientific importance. For novelty/originality, the correlation is somewhat weaker (−0.27 and 0.28, respectively), and it is even weaker for solidity (−0.18 and 0.18, respectively).

It is not surprising that the agreement is the poorest for solidity, considering that solidity per se is not the reason why a publication is cited in subsequent research (Aksnes et al., 2019). Rather, it may be a necessary but not sufficient criterion for a publication to be considered worth citing, at least according to a normative citation behavior model (see, for example, Bornmann & Daniel, 2008; Tahamtan & Bornmann, 2019). Similar reasoning can be provided for novelty/originality. A publication may report research that is original in approach, but if the results do not make interesting contributions to current knowledge, the publication may not be cited.

Although the association is statistically significant for all quality dimensions and both indicators analyzed, it is not strong. As a rule of thumb, correlation coefficients below 0.35 are considered weak (Taylor, 1990). In this study, only one case (scientific importance) has a higher value (just about): 0.35/0.37. This would correspond to a coefficient of determination of approximately 0.13 (r^2), meaning that 13% of the variance in the indicator can be "explained" by author ratings. Still, the characteristics of the response distribution must be taken into consideration in the interpretation. Generally, the value of the correlation coefficient will be larger when there is more variability among the observations than when there is less variability (Goodwin & Leech, 2006). As noted, a large majority of the papers were rated with scores 3–5, reducing the variability. In addition, certain range restrictions were applied in the initial identification of the researchers. This means that the identified correlation should be interpreted as stronger than the raw coefficients might suggest.

The results show that at the level of individual publications, citation metrics quite often do not correspond with author assessments. The correspondence is highest for highly cited publications. Many of the less-cited publications are considered to be of good quality and obtain high scores on the quality dimensions. This supports the conclusion that citations are unreliable at the level of individual articles but have stronger reliability at aggregated levels. A similar point is made by Traag and Waltman (2019), who claim that when analyzed at the level of articles, the strength of the correlation is much weaker than when addressed for aggregated units.

How do our results compare with previous studies comparing peer ratings and citation indicators for individual publications? A relevant example is the study of how the REF 2014 quality scores correlated with metrics (HEFCE, 2015), reporting a Spearman's rank correlation coefficient of −0.29 for the percentile indicator and 0.28 for an MNCS-like indicator (field-weighted citation impact). These results are lower than the results of our study on scientific importance but higher than those obtained for the other quality dimensions. The study by Baccini et al. (2020) also showed an overall degree of agreement consistent with the results of the HEFCE report. We also find other studies addressing this at the level of individual papers, reporting correlation coefficients around 0.2 (Borchardt et al., 2018; Patterson & Harris, 2009).

The conclusions of the HEFCE report have, however, been criticized by Traag and Waltman (2019). They noted that previous studies showed much higher agreement between metrics and

peer reviews in the REF. In their view, the issue should be addressed at an aggregated level (institutional) rather than at the level of individual papers. The reason for this is that the goal of the REF is not to assess individual publications but to assess institutional research. Using an alternative approach, Traag and Waltman (2019) were able to show much stronger correlations than those found in the HEFCE report. The fact that citations are not accurate at the individual publication level does not hinder satisfactory accuracy at higher levels.

4.3. The Results Are Consistent Across Fields

As a subordinate issue to the present study, we aimed to address field differences in the correspondence between self-perceived quality dimensions and citation scores. Here, we did not observe large differences across domains. In comparison, the HEFCE (2015) study found large variations across fields, the strongest correlation for Clinical Medicine (Spearman's rank correlation of 0.64 for the FWCI) and less than 0.2 in several Social Science disciplines and the Humanities (SSH). As noted in the introduction, a common view is that citation indicators have less reliability in SSH. Our study of author perceptions does not support the notion that SSH should be more problematic in this respect, although the problem of limited coverage will be more severe. It should be noted that some of the questions (e.g., concerning solidity) mainly refer to empirical sciences. It is therefore an open question how this is interpreted within the Humanities.

5. CONCLUSIONS

This study has shown that citation metrics of publications covary with the authors' own assessments of their quality dimensions. The association is statistically significant, although not strong. At aggregated levels, there is a distinct pattern in which rankings decline with declining citation metrics. Generally, the highest accuracy is obtained for highly cited publications, which are usually considered to have high research quality attributes. In terms of policy implications, this means that citations are not reliable indicators at the level of individual articles, while at aggregated levels, the validity is higher, at least according to how authors perceive quality. Hence, it is important to take the level of aggregation into account when using citations as performance measures. Despite statistically significant covariations for all quality dimensions analyzed, the association is strongest for scientific importance.

ACKNOWLEDGMENTS

We are thankful to the R-QUEST team for their input and comments on the questionnaire and a previous draft of the paper. We would also like to thank three anonymous reviewers for their valuable comments on earlier drafts of the manuscript. Last but not least we would like to thank the many researchers who took the time to fill in the questionnaire.

AUTHOR CONTRIBUTIONS

Dag W. Aksnes: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Writing—original draft, Writing—review & editing. Lone Wanderås Fossum: Data curation, Investigation, Software, Writing—review & editing. Fredrik Niclas Piro: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This research was funded by the Norges Forskningsråd under grant number 256223 (the R-QUEST Centre).

DATA AVAILABILITY

Data are not available. The participants of this study did not give written consent for their data to be shared publicly. Bibliographic record data cannot be released due to copyright/license restrictions.

REFERENCES

Aksnes, D. W. (2006). Citation rates and perceptions of scientific contribution. *Journal of the American Society for Information Science and Technology*, *57*(2), 169–185. https://doi.org/10.1002/asi.20262

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 1–17. https://doi.org/10.1177/2158244019829575

Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy*, *38*(6), 895–905. https://doi.org/10.1016/j.respol.2009.02.001

Aksnes, D. W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science*, *4*(1), 1–21. https://doi.org/10.2478/jdis-2019-0001

Baccini, A., Barabesi, L., & De Nicolao, G. (2020). On the agreement between bibliometrics and peer review: Evidence from the Italian research assessment exercises. *PLOS ONE*, *15*(11), e0242520. https://doi.org/10.1371/journal.pone.0242520, PubMed: 33206715

Bianco, M., Gras, N., & Sutz, J. (2016). Academic evaluation: Universal instrument? Tool for development? *Minerva*, *54*(4), 399–421. https://doi.org/10.1007/s11024-016-9306-9

Borchardt, R., Moran, C., Cantrill, S., Chemjobber, Oh, S. A., & Hartings, M. R. (2018). Perception of the importance of chemistry research papers and comparison to citation rates. *PLOS ONE*, *13*(3), e0194903. https://doi.org/10.1371/journal.pone.0194903, PubMed: 29590216

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Caon, M., Trapp, J., & Baldock, C. (2020). Citations are a good way to determine the quality of research. *Physical and Engineering Sciences in Medicine*, *43*(4), 1145–1148. https://doi.org/10.1007/s13246-020-00941-9, PubMed: 33165822

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, *51*(7), 635–645. https://doi.org/10.1002/(SICI)1097-4571(2000)51:7<635::AID-ASI6>3.0.CO;2-H

Cole, J., & Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the *Science Citation Index*. *American Sociologist*, *6*, 23–29.

Coupe, T. (2013). Peer review versus citations—An analysis of best paper prizes. *Research Policy*, *42*(1), 295–301. https://doi.org/10.1016/j.respol.2012.05.004

Dirk, L. (1999). A measure of originality: The elements of science. *Social Studies of Science*, *29*(5), 765–776. https://doi.org/10.1177/030631299029005004

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, *113*(1), 219–236. https://doi.org/10.1007/s11192-017-2483-y

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *Journal of Experimental Education*, *74*(3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266

Harzing, A.-W. (2018). Running the REF on a rainy Sunday afternoon: Can we exchange peer review for metrics? In R. Costas, T. Franssen, & A. Yegros-Yegros (Eds.), *Proceedings of the 23rd International Conference on Science and Technology Indicators* (pp. 339–345). Centre for Science and Technology Studies (CWTS), Leiden.

HEFCE. (2015). *The metric tide: Correlation analysis of REF2014 scores and metrics (Supplementary report II to the independent review of the role of metrics in research assessment and management)*. https://responsiblemetrics.org/the-metric-tide/

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a, PubMed: 25903611

Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A., & Klavans, R. (2014). Bibliometrics: Is your most cited work your best? *Nature*, *514*(7524), 561–562. https://doi.org/10.1038/514561a, PubMed: 25355346

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121, PubMed: 10626367

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674054158

Langfeldt, L., Nedeva, M., Sorlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. *Minerva*, *58*(1), 115–137. https://doi.org/10.1007/s11024-019-09385-2

Langfeldt, L., Reymert, I., & Aksnes, D. W. (2021). The role of metrics in peer assessments. *Research Evaluation*, *30*(1), 112–126. https://doi.org/10.1093/reseval/rvaa032

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Marx, W., & Bornmann, L. (2015). On the causes of subject-specific citation rates in Web of Science. *Scientometrics*, *102*(2), 1823–1827. https://doi.org/10.1007/s11192-014-1499-9

Mendoza, M. (2021). Differences in citation patterns across areas, article types and age groups of researchers. *Publications*, *9*(4), 47. https://doi.org/10.3390/publications9040047

Miranda, R., & Garcia-Carpintero, E. (2018). Overcitation and overrepresentation of review papers in the most cited papers. *Journal of Informetrics*, *12*(4), 1015–1030. https://doi.org/10.1016/j.joi.2018.08.006

Moed, H. F. (2005). *Citation analysis in research evaluation*. Berlin: Springer.

Nygaard, L. P., Aksnes, D. W., & Piro, F. N. (2022). Identifying gender disparities in research performance: The importance of comparing apples with apples. *Higher Education*, *84*, 1127–1142. https://doi.org/10.1007/s10734-022-00820-0

Ochsner, M., Hug, S., & Galleron, I. (2017). The future of research assessment in the humanities: Bottom-up assessment procedures. *Palgrave Communications*, *3*, 17020. https://doi.org/10.1057/palcomms.2017.20

Patterson, M. S., & Harris, S. (2009). The relationship between reviewers' quality-scores and number of citations for papers published in the journal *Physics in Medicine and Biology* from 2003–2005. *Scientometrics*, *80*(2), 343–349. https://doi.org/10.1007/s11192-008-2064-1

Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva*, *1*, 54–73. https://doi.org/10.1007/BF01101453

Porter, A. L., Chubin, D. E., & Xiao-Yin, J. (1988). Citations and scientific progress: Comparing bibliometric measures with scientist judgments. *Scientometrics*, *13*(3–4), 103–124. https://doi.org/10.1007/BF02017178

Savov, P., Jatowt, A., & Nielek, R. (2020). Identifying breakthrough scientific papers. *Information Processing & Management*, *57*(2), 102168. https://doi.org/10.1016/j.ipm.2019.102168

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628–638. https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<628::AID-ASI5>3.0.CO;2-0

Shibayama, S., & Wang, J. (2020). Measuring originality in science. *Scientometrics*, *122*(1), 409–427. https://doi.org/10.1007/s11192-019-03263-0

Sivertsen, G. (2018). The Norwegian model in Norway. *Journal of Data and Information Science*, *3*(4), 3–19. https://doi.org/10.2478/jdis-2018-0017

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, *12*(2), 461–480. https://doi.org/10.1016/j.joi.2018.03.007

Smolinsky, L., Sage, D. S., Lercher, A. J., & Cao, A. (2021). Citations versus expert opinions: Citation analysis of featured reviews of the American Mathematical Society. *Scientometrics*, *126*(5), 3853–3870. https://doi.org/10.1007/s11192-021-03894-2

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

Taylor, R. (1990). Interpretation of the correlation-coefficient–A basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39. https://doi.org/10.1177/875647939000600106

Traag, V. A., & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, *5*, 29. https://doi.org/10.1057/s41599-019-0233-x

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, *6*(4), 700–711. https://doi.org/10.1016/j.joi.2012.07.008

Waltman, L., & Costas, R. (2014). F1000 Recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, *65*(3), 433–445. https://doi.org/10.1002/asi.23040

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*(2), 372–379. https://doi.org/10.1002/asi.22775

Waltman, L., & Van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*(4), 833–849. https://doi.org/10.1016/j.joi.2013.08.002

Weinberg, A. M. (1963). Criteria for scientific choice. *Minerva*, *1*(2), 159–171. https://doi.org/10.1007/BF01096248

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., ... Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. HEFCE. https://responsiblemetrics.org/the-metric-tide/. https://doi.org/10.4135/9781473978782