Noname manuscript No.
(will be inserted by the editor)

Do Citations and Readership Identify Seminal Publications?

Drahomira Herrmannova · Robert M. Patton · Petr Knoth · Christopher G. Stahl

Received: date / Accepted: date

Abstract This work presents a new approach for analysing the ability of existing research metrics to identify research which has strongly influenced future developments. More specifically, we focus on the ability of citation counts and Mendeley reader counts to distinguish between publications regarded as seminal and publications regarded as literature reviews by field experts. The main motivation behind our research is to gain a better understanding of whether and how well the existing research metrics relate to research quality. For this experiment we have created a new dataset which we call TrueImpactDataset and which contains two types of publications, seminal papers and literature reviews. Using the dataset, we conduct a set of experiments to study how citation and reader counts perform in distinguishing these publication types, following the intuition that causing a change in a field signifies research quality. Our research shows that citation counts work better than a random baseline (by a margin of 10%) in distinguishing important seminal research papers from literature reviews while Mendeley reader counts do not work better than the baseline.

First author (ORCID: 0000-0002-2730-1546)
The Open University, Walton Hall, Milton Keynes, UK
Phone: +1 (865) 253-9980
E-mail: drahomira.herrmannova@open.ac.uk
Current address: Oak Ridge National Laboratory, Oak Ridge, TN, USA

Second author (ORCID: 0000-0002-8101-0571)
Oak Ridge National Laboratory, Oak Ridge, TN, USA

Third author (ORCID: 0000-0003-1161-7359)
The Open University, Walton Hall, Milton Keynes, UK

Fourth author (ORCID: 0000-0002-2070-1555)
Oak Ridge National Laboratory, Oak Ridge, TN, USA

Keywords Information Retrieval, Scholarly Communication, Publication Datasets, Data Mining, Research Evaluation, Bibliometrics, Altmetrics

1 Introduction

“Quality” is a commonly used term in research evaluation. It has been stated that the goal of peer review is ensuring only high-quality research gets published [Kelly et al., 2014], and that the focus of evaluative scientometrics has recently been directed at the quality of research publications [Bornmann and Haunschild, 2017]. However, what exactly is research quality? In scientometrics, research impact, which has been seen as a proxy to quality, has typically been measured in terms of the number of citations [Butler, 2008, Abramo et al., 2010, Bornmann and Haunschild, 2017], nevertheless, many researchers have pointed out issues associated with making such a connection [Meho, 2007, Adler and Harzing, 2009, Adler et al., 2009, MacRoberts and MacRoberts, 2010b, Onodera and Yoshikane, 2015, Ricker, 2017]. The reasons why the connection between citation counts and quality are considered problematic are many, from the fact citations may be used to criticise as well as praise [Onodera and Yoshikane, 2015] to the fact that quality is a complex and multifaceted concept which cannot easily be expressed in a single indicator [Ricker, 2017]. Peer review, especially when it comes to journals with high impact factor, is often considered to be the best available measure of quality [Garfield, 2003, Bornmann and Daniel, 2005, Kreiman and Maunsell, 2011]; however, this method of recognising high quality research also has its drawbacks, including reviewer bias [Teixeira da Silva and Dobránszki, 2015] and high disagreement in peer review decisions [Francois, 2015].

When asked about research quality, scientific impact and excellence, most people usually refer to the volume of change produced in a particular field (contribution to research, how much did a piece of work move the field forward), rather than referring to the educational (or other types of) impact generated [Sternberg and Gordeeva, 1996]. Similarly, when reviewing journal publications, the most important factor influencing the reviewers’ decision to accept or reject the paper is its perceived research contribution [Bornmann et al., 2008, Nedić and Dekanski, 2016]. This is also the case for many national evaluation systems [Research Excellence Framework, 2012, Tertiary Education Commission, 2013, Australian Research Council, 2015]. Therefore, in this paper we study how well the existing metrics, particularly citation counts, work in distinguishing publications that generate a very high amount of research contribution from publications that do not. The main motivation behind our research is to gain a better understanding of whether and how well the existing research metrics relate to research quality, however, we believe our study will also prove useful in testing new research metrics.

We use seminal publications and literature reviews as characteristic examples of publications generating very high and very low volume of change. Indeed, the definition of the word *seminal* according to the Oxford Dictionary is “strongly influencing later developments” while the definition of the word *review* is “a report on or evaluation of a subject or past events”, which matches our understanding of the difference between these two types of papers. Hence, if one of the goals of research evaluation is recognising publications which contributed significantly to their field, seminal papers should perform better under such evaluation than literature reviews, which by definition do not generate a change in the field[1].

Therefore, we study how well the existing metrics discriminate between these two types of papers. Our results show that existing metrics help in distinguishing between seminal publications and literature reviews, albeit with room for improvement. We believe this is an important finding demonstrating more attention may need to be paid to publication type in research evaluation, especially as these two types of papers are weighted equally when used in research evaluation metrics, such as in JIF [McVeigh and Mann, 2009] and the h-index, although literature reviews are sometimes excluded from research evaluation studies, such as in the Research Excellence Framework. The work presented in this paper is conducted on a new dataset of seminal publications and literature reviews which we call TrueImpactDataset and which was built for this study. We share this dataset with the research community[2] to help the development of new research evaluation metrics.

This paper is organized as follows. First, in Section 2 we present the related work. In Section 3, we describe our research question and how we aim to answer it. In Section 4, we explain how the dataset was created. Section 5 presents some statistics of the dataset and Section 6 the results of the experiment in which we examine the value of citations and Mendeley reader counts in predicting the type of a paper. In Section 7, we discuss our findings and the properties an ideal dataset for evaluating research metrics should have.

2 Related work

The suitability of current metrics for assessing the value of research outputs has been studied extensively in the literature, especially the suitability of citations, however, other indicators [Bornmann, 2014, Thelwall and Kousha, 2015a], including Mendeley readership [Bornmann, 2015, Thelwall and Kousha, 2015b], have been studied as well.

The existing studies have typically approached the question using one of the following two methods: **1) by studying the unit of measurement itself**, for instance in the case of citations by studying the motivations of scientists for choosing to reference or to not reference specific papers [Harwood, 2009, MacRoberts and MacRoberts, 2010a] (a review of studies on citing behaviour is available in [Bornmann and Daniel, 2008]), the characteristics of citations, such as the placement [Bertin et al., 2016], and the context [Hu et al., 2015]

[1] With some exceptions, notably systematic reviews, which are a key practice in evidence-based medicine

[2] `http://trueimpactdataset.semantometrics.org/`

of citations in text, or in the case of Mendeley readership the reasons for bookmarking specific papers [Mohammadi et al., 2016]; or **2) by studying what a given metric represents**, for example by analysing the characteristics of highly cited papers [Aksnes, 2003, Antonakis et al., 2014, Van Noorden et al., 2014] or by comparing the data with another metric [Bornmann and Leydesdorff, 2015, Bornmann and Haunschild, 2015], typically by performing a correlation analysis.

Our work has both similarities and dissimilarities with the studies mentioned in the previous paragraph. Similarly as the works studying highly cited publications, we analyse whether a high number of received citations reflects the shift a paper caused (or didn't cause) in its field. Interestingly, two of the mentioned studies have found a high proportion of the top cited papers to be literature reviews [Aksnes, 2003] or method and software descriptions [Van Noorden et al., 2014]. In contrast to previous work, we concentrate on analysing how well important seminal papers perform under current evaluation methods in comparison to literature reviews, rather than focusing on characterising highly cited papers, or understanding what the existing metrics measure.

Our work is also close to several recent efforts [Teufel et al., 2006, Wan and Liu, 2014, Zhu et al., 2015, Valenzuela et al., 2015, Pride and Knoth, 2017] in which the authors argue that not all citations are equal and that identifying which citations are important is necessary for better understanding of published research. Our work provides quantitative evidence further motivating this strand of research, as we show that while using citations works to some extent for distinguishing seminal publications research from literature reviews, there is a room for improvement. As a future work we would like to test the models presented in these studies on our dataset to see whether classifying citations according to their importance will help distinguish seminal publications from literature reviews.

3 Methodology

This paper aims to answer the following research question: “How well do citation and reader counts distinguish important seminal publications from literature reviews?” To answer this question we adopt the following method.

As mentioned in the introduction, when talking about evaluation of research outputs, an important dimension is the volume of change produced in a research area (how much was the area pushed forward thanks to a given piece of work) [Sternberg and Gordeeva, 1996, Bornmann et al., 2008, Research Excellence Framework, 2012, Tertiary Education Commission, 2013, Australian Research Council, 2015]. This amount of change has been discussed and studied from different perspectives [Yan et al., 2012, Knoth and Herrmannova, 2014, Whalen et al., 2015, Valenzuela et al., 2015, Patton et al., 2016]. We were looking for a sample of research publications representing such work and we believe seminal research papers constitute such sample. To provide a clear comparison we were also interested in review publications (papers presenting a survey of a research area). While these papers are often highly cited [Seglen, 1997, Aksnes, 2003] they usually don't present new original ideas. In this paper we study how well citation counts and Mendeley reader counts distinguish between these two types of papers.

To our knowledge, there currently isn't any dataset which would categorize papers into these two categories. We were therefore left with the task of creating such dataset ourselves. We have designed an online survey to collect the dataset. The format of the survey, the number of collected responses and other details are presented in Section 4.1. In the following section (5) we analyze the dataset to understand whether it is suitable for our purposes.

In order to answer our research question, we have designed a simple experiment. We chose citation counts and Mendeley readership as representatives of bibliometrics and altmetrics, as these two measures are both well known and are being used as measures of impact of published research in many settings [REF 2014, 2012, Wilsdon et al., 2015]. We then classify the papers in the collected dataset into two classes (seminal, review) using two models, a model using the papers' citation counts and a model using their Mendeley readership (Section 6).

4 Dataset creation

This section describes the dataset and the process used to create it. The dataset is publicly available for download[3].

4.1 Initial data collection

The goal was to create a collection of research publications consisting of two types of papers, seminal works, and literature reviews. We have used an online form to collect the references, which was composed of two sets of questions – questions about the respondent's academic background (their discipline, seniority and publication record) and questions which asked for a reference to a seminal paper and to a literature review, both related to the respondent's discipline. We have used the latest Research Excellence Framework (REF) units of assessment [Research Excellence Framework, 2014] as a list of disciplines when asking about the respondent's academic background because UK researchers are familiar with this classification.

The survey was sent to academic staff and research students from all faculties of the Open University (to 1,415 people in total). The reason why we contacted Open University researchers is because research at the Open University covers many disciplines, and because it is the largest university in the United Kingdom. We were therefore able to get a significant sample spanning multiple disciplines. Within three months we have received 184 responses (172 references to seminal papers and 157 to review papers), which represents a

[3] http://trueimpactdataset.semantometrics.org

13% response rate. The survey questions and email invitation are available online together with the dataset[3]. After removing empty and unidentifiable responses, we were left with 171 responses providing us with 166 seminal and 148 literature reviews.

4.2 Additional metadata

Once the survey was closed we have manually processed the data and collected the following information (by querying a search engine for the paper title and looking for a relevant page): a DOI, or a URL for papers for which we did not find a DOI, title, list of authors, year of publication, number of citations in Google Scholar and abstract. Where we had access to the full text, we have also downloaded the PDF. We were able to download 275 PDFs and 296 abstracts. Due to copyright restrictions, the PDFs are not part of the shared dataset[4]. This collection process took a single person several hours a day for about a week.

To obtain readership data, we have used the DOIs, or title and year of publication for papers without a DOI, to query the Mendeley API[5]. We were mainly interested in the number of readers of each paper. The dataset contains a snapshot of the Mendeley metadata we were working with. We were able to find 141 out of the 166 seminal papers and 125 out of the 148 literature reviews in Mendeley.

Furthermore, using the Web of Science (WoS) API[6] we managed to retrieve additional information for the seminal and literature review papers indexed by WoS. We queried the WoS API using publication DOIs. In this case we were mainly interested in the number of citing papers and cited papers (references).

5 Dataset analysis

To ensure the collected dataset is suitable for our task, we looked several statistics describing the dataset including statistics of publication age, distribution across disciplines and citation and readership statistics.

5.1 Size

The size of the dataset is presented in Table 1. The row *DOIs* shows the number of papers in the dataset for which we were able to find a DOI and the row *Seminal/review/total in WoS* shows how many of these DOIs appear in

[4] As there are Copyright Exceptions for text and data mining in some countries, such as in the UK, we are happy to provide the PDF documents for these purposes to researchers residing in these jurisdictions upon request.

[5] `http://dev.mendeley.com`

[6] `http://ipscience-help.thomsonreuters.com/wosWebServicesLite/WebServicesLiteOverviewGroup/Introduction.html`

the Web of Science database. The number of additional references which we collected using the WoS API is shown in the row *Citing & cited references*. The row *Total/unique authors* shows the total number of authors of all papers in the dataset and the number of unique author names. To count the unique names, we have compared the surname and all first name initials, in case of a match we consider the names to be the same (e.g. J. Adam Smith and John A. Smith will be counted as one unique name). The number of unique author names doesn't show the number of disambiguated authors, but gives us an indication of how many of the author names repeat in the dataset.

Responses	171
Seminal/review/total papers	166/148/314
Seminal/review/total in Mendeley	141/125/266
Seminal/review/total in WoS	51/59/110
DOIs	256
Abstracts	296
Total/unique authors	1334/1235
Citing & cited references	19,401

Table 1 Size of the dataset used in the study. The field *Responses* refers to how many responses we received in the data collection survey and the field *Seminal/review/total papers* refers to how many papers of each type the responses yielded.

5.2 Publication age

Figure 1 shows a histogram of years of publication with literature reviews and seminal papers being distinguished by colour. Seminal papers in the dataset are on average about 9 years older than review papers. This shows literature reviews might age faster than seminal papers, which is consistent with our expectations. An explanation for this could be that literature reviews theoretically become outdated as soon as the first new piece of work is published after the publication of the review. Because the seminal papers are on average older this also means these papers had more time to attract citations. This is another reason to expect seminal papers to be distinguishable from literature reviews by citations and readership as features. Descriptive statistics of years of publication both sets are presented in Table 2.

5.3 Disciplines

Figure 2 shows a histogram of papers per discipline. We have used the information we got about the respondents' academic background to assign papers to

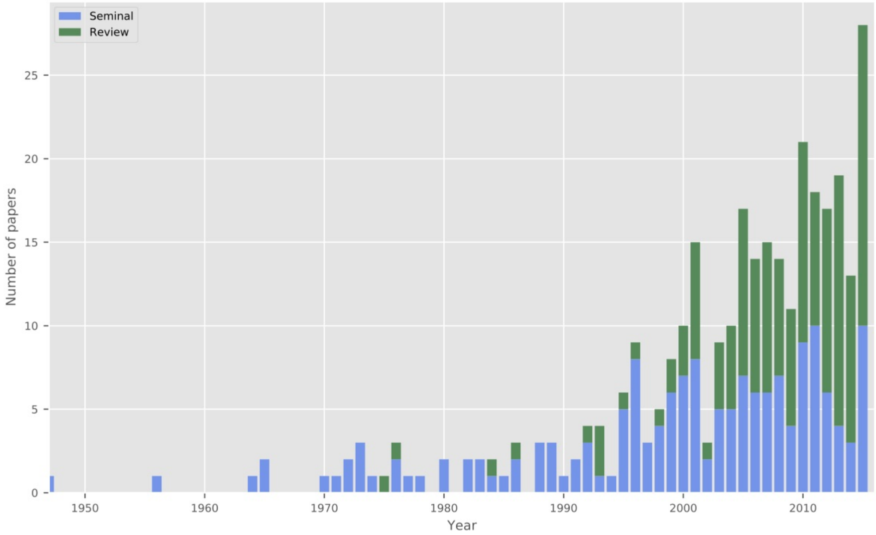

Fig. 1 Histogram of publication years. Seminal publications in the dataset are represented as blue bars and literature reviews as green bars.

	Seminal	**Review**	**Overall**
Count	166	148	314
Mean	1999	2008	2003
Median	2002	2010	2006
Min	1947	1975	1947
Max	2016	2016	2016

Table 2 Descriptive statistics of publication age for both types of papers. The row *Count* shows how many publications are included in each column.

disciplines. The distribution of papers per discipline is to a certain degree consistent with other studies, which have reported Computer Science and Physics to be among the larger disciplines in terms of number of publications, however, Medicine and Biology are typically reported to be the most productive [Althouse et al., 2009, D'Angelo and Abramo, 2015]. The distribution is therefore probably more representative of size of faculties of the Open University than of productivity of scientific disciplines in general, however, we believe this does not influence our study.

When answering the questions about academic background, 22 respondents have selected "Other" instead of one of the listed disciplines, these 22 responses provided us with 40 papers in total. We looked at the detailed description of their topic provided by the respondents. 9 of them are related to astronomy (the descriptions provided were "Binary stars", "Martian meteorites", "cosmochemistry", "Planetary sciences" and "planetology"), 4 could be classified

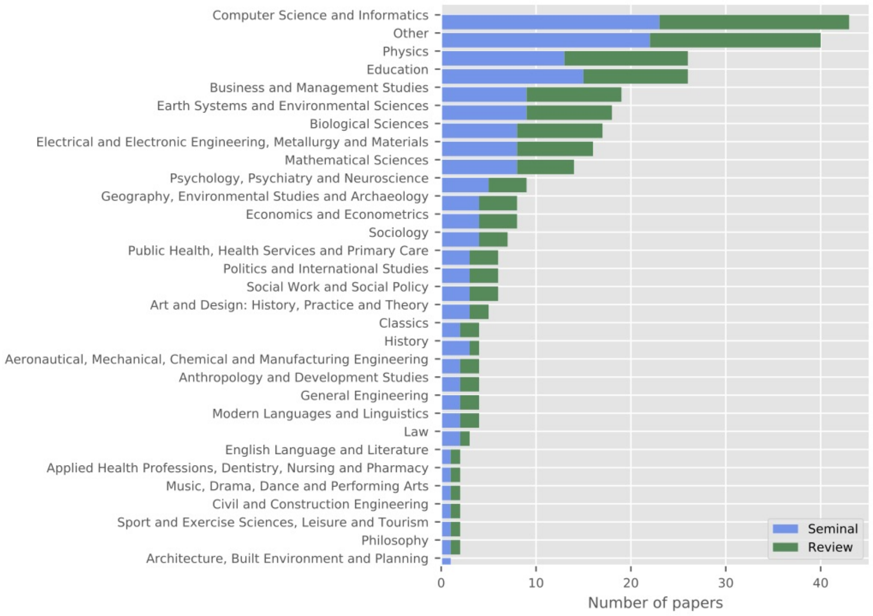

Fig. 2 Histogram of publication disciplines. In the histogram, seminal publications are represented as blue bars and literature reviews as green bars.

as computer science (“virtual reality” and “Natural Language Understanding, Spoken Language Understanding”), the rest relate to different areas (e.g. “Microbial degradation of plastic” or “MOOC”).

5.4 Citations and readership

The dataset contains two basic measures related to publication impact and utility – citation counts, which we manually collected from Google Scholar, and the number of readers in Mendeley. We also had access to the number of citations in Web of Science and while we couldn’t make these data available together with the dataset, we provide an analysis of the WoS citations and a comparison with the other two metrics.

Table 3 shows basic statistics of Google Scholar citation counts and Mendeley readership of each paper in the dataset. We consider the readership of papers which we didn’t find in Mendeley to be 0 (as papers are added to the Mendeley database by their readers). It is interesting to notice that while seminal papers in our dataset are on average cited more than review papers, this is not the case for readership, in fact literature reviews attract more readers than seminal papers despite being on average younger (Section 5.2). We believe this is an important finding as readership counts are being more and more frequently used as a measure of impact complementary to citations [Piwowar and Priem, 2013, Maflahi and Thelwall, 2016, Priem, 2014]. We believe the fact that literature reviews are more read than seminal papers, while being less cited, suggests that readership can be perceived more as a measure of popularity than importance.

	Google Scholar citations			**Mendeley readership**		
	Seminal	**Review**	**Overall**	**Seminal**	**Review**	**Overall**
Mean	2,458	519	1,544	240	368	306
Median	249	109	194	45	42	46
Std	8,885	1,197	6,575	894	1,566	1,264
Min	0	0	0	0	0	0
Max	85,376	12,099	85,376	10,258	15,516	15,597

Table 3 Descriptive statistics of Google Scholar citation counts and of Mendeley readership. Each column presents statistics for the selected type of paper (columns *Seminal* and *Review*) or for all papers regardless of type (column *Overall*).

Table 4 shows a comparison of the number citations obtained from Google Scholar and from WoS. This table includes only those 110 papers (51 seminal and 59 survey papers) which appear in WoS. The higher citation numbers coming from Google Scholar are not surprising as Google Scholar's wider coverage of academic outputs is well known [Harzing and Alakangas, 2016a, Harzing, 2016]. This wider coverage is also demonstrated by the fact that we were able to find only 110 out of the 314 papers used in our study in WoS.

	Google Scholar			**Web of Science**		
	Seminal	**Review**	**Overall**	**Seminal**	**Review**	**Overall**
Mean	814	429	607	523	255	379
Median	211	216	214	144	94	105
Std	1,599	566	1,175	926	373	697
Min	2	0	0	1	0	0
Max	8,246	2,446	8,246	4,753	1,709	4,753

Table 4 Descriptive statistics of citation counts acquired from Google Scholar and Web of Science. This table includes only those 110 papers, which we were able to find in WoS. Each column presents statistics for the selected type of paper (columns *Seminal* and *Review*) or for all papers regardless of type (column *Overall*).

This low coverage provided by Web of Science can be seen as a problem, especially given the fact WoS misses some key seminal papers and overall misses more seminal papers than literature reviews. For example, a recent publication by Krizhevsky et. al. [Krizhevsky et al., 2012], a seminal deep learning paper which has caused a shift in the area of artificial intelligence/computer vision, is missing in WoS, but has (at the time of writing this paper) attracted almost 8000 citations in GS since its publication in 2012. This problem isn't limited to WoS either, Scopus for example also does not index the publication, and while Mendeley does, most of the associated meta-data is inaccurate. The most probable reason for these exclusions is that the conference proceedings for this paper are not published through a major publisher but instead by the conference itself and self-hosted on their website. We believe this is an interesting point as it shows important seminal work isn't always published by the traditional routes of journals or known publishers. With the recent changes in scholarly communication towards Open Access, Open Science, Arxiv, self hosting, etc. the very definition of "published" no longer has a universal standard and we believe it is reasonable to expect that this will continue with higher frequency as the communities continue to change over time.

In order to compare whether the two databases rank papers similarly we have correlated the citation counts (see Table 5). Both correlations are weaker for seminal papers, however this could be caused by the age difference between the two types of papers as the databases might have a lower coverage of older publications. On the other hand, the overall correlations may be slightly inflated due to older articles having on average higher citation values. This might, in combination with seminal publications in our dataset being older, cause ranks produced using citation counts to be more similar for the two databases and as a consequence may increase the correlations. One solution to overcome this limitation would be correlating only articles published in the same year. However, studying the differences between the databases is not the main aim of our study and for this reason we do not present here the separate correlations. Overall, both Pearson and Spearman correlations are otherwise strong. We believe the strong correlations show using citation data from these two databases will produce similar results.

	Spearman	**Pearson**
Seminal	$0.8581, p \ll .001$	$0.6775, p \ll .001$
Review	$0.9696, p \ll .001$	$0.9588, p \ll .001$
Overall	$0.9281, p \ll .001$	$0.7254, p \ll .001$

Table 5 Correlation between Google Scholar and Web of Science citation counts. Each field in the table represents Spearman or Pearson correlation between citation counts found in Google Scholar and WoS. The correlation is calculated for seminal publications and literature reviews separately as well as for all papers combined together.

6 Experiment & Results

In this section, we present the results of the experiment, the aim of which was to test whether citation or readership counts work as a discriminating factor for distinguishing seminal papers and literature reviews. These two measures, and especially citation counts, are frequently used as proxies for scientific influence and quality. For example, citation counts are the basis for calculating JIF, where the calculation doesn't take into account the differences between types of research papers (pure research papers and literature reviews are both used as input with equal weight) [Thomson Reuters]. Amount of research contribution is often indicated as an important dimension of research quality [Sternberg and Gordeeva, 1996, Bornmann et al., 2008, Nedić and Dekanski, 2016, Research Excellence Framework, 2012, Tertiary Education Commission, 2013, Australian Research Council, 2015]. Thus, we study how well these two measures distinguish between seminal publications and literature reviews, which respectively represent publications generating very high and very low amounts of research contribution.

In order to test our hypothesis we use these two metrics to classify the papers into the two classes (seminal, review) – i.e. we use citation and reader counts as two features for creating classification models. As a baseline we use a model which classifies all papers as seminal, as that is the majority class. This baseline model achieves the accuracy of 53%. We calculate accuracy as the proportion of correctly classified publications, or more formally:

$$acc = \frac{TP + TN}{N} \tag{1}$$

where the category *seminal* is our positive class, TP (true positives) is the number of items correctly labelled as belonging to the positive class, TN (true negatives) is the number of items correctly labelled as not belonging to the positive class, and N is the number of all items (publications).

Before running the experiments we first perform a statistical test to see whether the citation/readership distributions of seminal and review papers differ. We perform a one-tailed independent t-test with the null hypothesis stating that the means of the two groups are equal. The results we get are $p = 0.0063$ for citations and $p = 0.1666$ for reader counts. In case of citations, for a significance threshold of 1% we reject the null hypothesis. Because we know the mean number of citations of the seminal papers is higher (Table 3), we conclude seminal papers are cited significantly more than literature reviews. In case of readership, we accept the null hypothesis that the distributions of reader counts of seminal and review papers are the same (that is the number of readers doesn't distinguish between the two groups). However, one possible explanation for this is the typical shape of the distribution of Mendeley readers. It has been shown that older articles tend to have less readers in Mendeley [Thelwall and Sud, 2016]. In combination with seminal publications in our dataset being older, this could lead to seminal publications being indistinguishable in terms of reader counts from literature reviews. For this reason we do not remove Mendeley reader counts from further analysis. To better understand how well each metric works in distinguishing between the two groups, we use citations and readership as features in a classification experiment.

6.1 Experimental setup and choice of a model

The classification experiment relies on two approaches. First, we use a leave-one-out cross-validation setup, that is we repeatedly train on all but one publication and then test the performance of the model on the publication we left out of the training. We do this for all publications in the set. However, because in some cases, due to the size of the dataset, leaving out even one publication can affect the performance of the model, we also find the performance of the ideal model, that is we train the model on all available data (all publications in the dataset) without leaving any publication out. This gives us an upper bound of performance on our dataset.

We run three separate experiments. First, we train and test our models on all publications, regardless of their age or discipline. This gives us an idea of how well do both metrics perform across disciplines and regardless of time. We call this the aggregate model (Section 6.2). Next, we split the data by discipline and create separate models for each discipline (Section 6.3). The reason for this is that both citation patterns as well as Mendeley coverage and usage patterns tend to differ across disciplines [Mohammadi et al., 2015, Waltman, 2016]. Finally, we split the data by publication years and create separate models for each year (Section 6.4). There are two reasons why splitting the publications by year is important. Firstly, when it comes to citation counts, older publications have more time to accumulate citations than newer publications and will therefore be often cited higher than newer publications [Waltman, 2016]. Because the seminal publications in our dataset are on average older than the literature reviews, this could significantly influence our results. Secondly, when it comes to Mendeley reader counts, their distribution tends to have different shape than the distribution of citation counts. In particular, newer publications often have more readers [Thelwall and Sud, 2016]. It would be interesting to also split the data by both discipline and year, however, we weren't able to do this due to the size of the dataset, as the resulting groups would be too small for analysis. In all three cases we also create two separate models – one model trained using citation counts as the only feature and one model trained using Mendeley reader counts as the only feature.

The model we use to classify papers based on their citation or reader counts works in the following way: if the total number of citations (or the number of readers) for a given paper is equal to or greater than a selected threshold we classify the paper as seminal, otherwise as a literature review. To do this, we use the threshold which achieves the best accuracy on the training data. We find this threshold by calculating the accuracy for all thresholds in the interval $[0, max(citation\_count)]$ for the model using citation counts and $[0, max(reader\_count)]$ for the model using reader counts. If there is more than one such threshold, we use the average value of all best thresholds. For the ideal model we choose any of the best thresholds, as all will have the same performance. The reason why we chose the this simple model instead of a machine learning model such as SVM or Naïve Bayes is that our model better reflects how research metrics are used by scientists in decision making.

6.2 Aggregate model

Table 6 shows the confusion matrix for the leave-one-out cross-validation scenario using a citation count threshold and not separating publications by discipline or age. This setup achieves an overall accuracy of 63%, which represents a 10% improvement over the baseline. All but two of the models trained in the cross-validation setup chose 51 citations as an optimal threshold (the two other thresholds were 52.4 and 52.5). The ideal model (trained on all available publications) achieves the accuracy of 63%. Table 7 shows the confusion matrix obtained by using reader counts as a feature. This model achieves an overall accuracy of 43%, which is 10% worse than the baseline. Most of the models (277) trained in the cross-validation setup chose 0 readers as the optimal threshold. The remaining models (37) chose 2.5 readers as a threshold. The performance of the ideal model is 53%, which is equal to the baseline.

		Predicted		
		Review	Seminal	Total
Actual	Review	19% (61)	28% (87)	148
	Seminal	9% (29)	44% (137)	166
	Total	90	224	314

Table 6 Confusion matrix for predicting the class of the paper using Google Scholar citation counts. In this case publications were separated into two groups using a citation counts threshold.

		Predicted		
		Review	Seminal	Total
Actual	Review	0% (0)	47% (148)	148
	Seminal	10% (32)	43% (134)	166
	Total	32	282	314

Table 7 Confusion matrix for predicting the class of the paper using Mendeley reader counts. To produce this matrix, publications were separated into two groups using a Mendeley reader counts threshold.

6.3 Discipline based model

This model uses discipline information to first split the papers into groups. For all separate groups we then perform the same statistical test and classification experiment using both citation and reader counts. In this case, we remove all papers labeled as “Other”. Furthermore, we remove all subject areas which contain less than two of each type of papers, to be able to train and test the models on representatives of both seminal and review papers. The p-value is greater than 1% for all remaining disciplines and for both citation and reader counts, which means in all cases we accept the null hypothesis of equal averages. All p-values are shown in Appendix A, Table 13.

The overall cross-validation accuracy is 45% for citations and 42% for reader counts, which is worse than the baseline (53%) in both cases. We believe this is due to the fact the baseline isn't dependent on the size of the data, while in the leave-one-out cross-validation, removing even one paper can change the performance of the model. Furthermore, the baseline method "knows" which class is the majority class, while our model doesn't use this information. Both of these factors make it harder to outperform the baseline. The results for separate disciplines are reported in Tables 14 and 15.

To calculate the overall accuracy, rather than counting average accuracy across all disciplines, we sum all confusion matrices and calculate the accuracy from the sum (Tables 8 and 9, this method is sometimes referred to as micro-averaging). The accuracy of the optimal model goes up in both cases, to 68% in the case of citations and to 63% in the case of readership. This shows that separating papers by discipline has the potential of improving the results.

		Predicted		
		Review	Seminal	Total
Actual	Review	24% (62)	24% (60)	122
	Seminal	31% (79)	21% (53)	132
	Total	141	113	254

Table 8 Overall classification results obtained from running the classification for each discipline separately. In this case, a citation counts threshold was used to classify publications as seminal or review. One model was created per discipline and the overall performance was obtained by summing confusion matrices from all disciplines and calculating accuracy from the sum.

		Predicted		
		Review	Seminal	Total
Actual	Review	17% (44)	31% (78)	122
	Seminal	27% (69)	25% (63)	132
	Total	113	141	254

Table 9 Overall classification results obtained from running the classification for each discipline separately. In this case, a Mendeley reader counts threshold was used to classify publications as seminal or review.

6.4 Year based model

We perform a similar experiment as in case of disciplines also for publication years. The results for this experiment are shown in Tables 10 and 11.

In this case we split the publications in the dataset into groups by the the year in which they were published and again leave out those groups which

		Predicted		
		Review	Seminal	Total
Actual	Review	40% (95)	17% (41)	136
	Seminal	28% (66)	15% (37)	103
	Total	161	78	239

Table 10 Overall classification results obtained from running the classification for each year separately, using citations as a feature. Similarly as in the previous case, one model was trained for each year and the overall accuracy was obtained by summing the confusion matrices for all years.

		Predicted		
		Review	Seminal	Total
Actual	Review	38% (90)	19% (46)	136
	Seminal	30% (71)	13% (32)	103
	Total	161	78	239

Table 11 Overall classification results obtained from running the classification for each year separately, using reader counts as a feature.

don't contain at least two papers of each type. The p-value is greater than 1% for all publication years (16). The overall cross-validation accuracy is 55% (Table 10) for citation counts and 51% (Table 11) for reader counts, which in the case of citation counts is an improvement both over the baseline (53%) and over the previous model trained per discipline. The accuracy of the optimal model is 69% in the case of citations and 65% in the case of reader counts. The full results for this experiment are reported in Tables 17 and 18.

7 Discussion

7.1 Results

Table 12 shows a summary of classification results of all three models. Overall the year based model performs better than the discipline based model, however, particularly when it comes to citation counts, this might be due to the distribution of survey and seminal publications in our dataset – as we have shown in Table 2, seminal papers in our dataset are on average older than literature reviews, which makes the year based classification easier. In reality papers published in a given year will be distributed more evenly. The performance of the discipline based model should be more stable, as the distribution of seminal and survey papers across disciplines in our dataset is more even. When it comes to Mendeley readership, the year-based model outperforms the other two models by a margin of almost 10%. One possible explanation for this is that Mendeley is a relatively new service which was created in 2008, however, many publications in our dataset, especially seminal publications, are much older. Furthermore, newer publications (up to a certain threshold) tend to have more readers [Thelwall and Sud, 2016]. Separating the publications by year is therefore important for a fair comparison. We haven't performed a classification across both disciplines and years as due to their wide distribution we weren't able to find enough examples belonging to the same discipline and year. The aggregate model outperforms the two other models, however, we believe this might be due to the size of the dataset.

Model	**Data**	**Accuracy**
Baseline	Citations	53%
	Readership	53%
Aggregate	Citations	63%
	Readership	43%
Discipline based	Citations	45%
	Readership	42%
Year based	Citations	55%
	Readership	51%

Table 12 Summary of all results. Column *Accuracy* shows the accuracy obtained in the leave-one-out cross-validation scenario.

7.2 Contribution of our work and comparison with existing literature

We believe this study is novel in two ways. Firstly, our experiments show that citation counts help in distinguishing important seminal research from literature reviews with a degree of accuracy (63%, i.e. 10% over the random baseline), while Mendeley reader counts don't work better than a random baseline on this task and our dataset. There has been much discussion whether citation counts are appropriate for use in evaluation of research outputs [Wilsdon et al., 2015]. We have used a new approach to study this question. In addition, our contributions include the creation of a novel dataset of 314 seminal publications and literature reviews, which is publicly available. We believe this dataset will be useful in developing and evaluating new metrics.

A number of previous works have studied highly cited literature to analyse and understand which factors make publications highly cited [Aksnes, 2003, Antonakis et al., 2014, Van Noorden et al., 2014]. Our work complements this strand of research, as in our work we study whether a high number of received citations reflects the shift a paper caused (or didn't cause) in its field. Our results are similar to results reported by Aksnes, who observed that a significant portion of top cited articles are review articles (12% of articles in their dataset [Aksnes, 2003]). We have chosen a slightly different approach, as we did not focus only on highly cited articles, but instead studied publications perceived by scientists as seminal and compared these publications to literature reviews. Our results suggest that to a degree, citation counts can distinguish between publications which caused a shift in their field, i.e. seminal publications, and literature reviews. Our work also complements existing methods for automated classification of research citations [Teufel et al., 2006, Valenzuela et al., 2015, Pride and Knoth, 2017] as it demonstrates the need for these methods. An interesting future direction would be to test whether the existing automated citation classification models help to improve performance in our task, i.e. whether using only "important" citations [Valenzuela et al., 2015, Pride and Knoth, 2017] helps in distinguishing seminal publications from literature reviews.

7.3 Limitations of our dataset

One limitation of our study is that we rely on the respondents' understanding of seminal publications and literature reviews and the fact our dataset is limited to responses from a single university. To get a better understanding of the validity of the data, we have verified the correctness of the responses belonging to the Computer Science and Informatics subset (43 publications), as that is an area most familiar to us. To do this, we have reviewed the publication titles and abstracts. The labelling of this subset matches our understanding of seminal and review publications except in three cases, a paper "From data mining to knowledge discovery in databases" which was labelled as seminal and papers "Process algebra for synchronous communication" and "Unifying heterogeneous and distributed information about marine species through the top level ontology MarineTLO" which were both labelled as a literature review. For these three papers we would flip the labels. We haven't however read the full papers and so our disagreement with the respondents could be caused by not knowing the content of the papers and/or not being experts in those areas. Overall, 93% of the publications (40 out of 43) reviewed by us were assigned a correct label.

The dataset was created through a survey conducted at the Open University (OU) in the UK. The OU is the largest university in the UK and one of the largest universities in Europe and in the world, with centres across the UK and Europe. Contacting academics from the OU has enabled us to get a significant sample spanning multiple disciplines and levels of seniority. Considering this fact and our manual examination of the dataset, we do not believe the results of our study would be significantly different if conducted at a different university, particularly in the UK. However, to overcome this limitation, as future work we are planning to cross-reference the data to ensure the validity of the entire dataset.

7.4 Database coverage

In Sections 5.1 and 5.4 we have compared the coverage of Google Scholar, which we used to collect publication metadata, Mendeley and Web of Science. Of the 314 publications in our dataset (all of which were found in Google Scholar) 266 (85%) were found in Mendeley and 110 (35%) in WoS. The WoS coverage is consistent with findings reported by Harzing and Alakangas, who have shown it can vary significantly when compared to Google Scholar (between approximately 17 and 66% depending on discipline) [Harzing and Alakangas, 2016b].

We have also observed higher citation counts in Google Scholar compared to WoS (Section 5.4). It has been shown that higher citation counts in Google Scholar are often caused by "stray citations", i.e. duplicate records and/or citations from non-traditional research outputs such as books, blogs, etc. [Harzing and Alakangas, 2016b], rather than better coverage. Therefore, to better understand the differences between the two databases, in Section 5.4 we have reported correlations between citation counts found in Google Scholar and Web of Science. Both correlations were very strong for literature reviews showing that for newer articles both databases will produce similar results. The numbers were significantly lower for seminal publications which may be caused by lower coverage of older articles. These differences in coverage may influence our results, particularly if one type of publication is represented better in a given database than the other. As future work it would be interesting to compare correlations between these two databases across years and both publication types.

Mendeley has been shown to have fairly high coverage [Priem et al., 2012, Haustein et al., 2014, Thelwall and Kousha, 2015b], which was also the case in our study. The coverage of seminal publications and literature reviews in Mendeley was the same, 85% in the case of seminal publications (141 out of 166 seminal papers were found in Mendeley) and 84% in the case of literature reviews (125 out of 148 reviews were found in Mendeley). One possible issue could stem from the fact majority of Mendeley readers are often PhD students, graduate students or postdocs [Haustein and Larivière, 2014]. These users may prefer to bookmark a certain type of publications (e.g. literature reviews more often than seminal publications). As future work it would be interesting to compare readership patterns of the two publication types in our dataset.

8 Conclusions

In this paper, we have shown that citation counts work 10% better than the baseline in distinguishing important seminal publications from literature reviews, while Mendeley reader counts don't work better than the baseline. We have performed a set of experiments using citation and reader counts to classify papers into seminal and review categories and showed that citations distinguish between these two types of papers with low to moderate accuracy (highest accuracy achieved in all experiments was 63%, while our baseline model achieved 53%), while reader counts don't distinguish between them at all (highest accuracy 51%). We believe this shows that while citations work to some degree, additional methods, such automated methods for classifying important citations [Teufel et al., 2006, Valenzuela et al., 2015, Pride and Knoth, 2017], may be needed.

In addition to quantifying the success rate when using citation count to distinguish seminal publications from literature reviews, we also presented a novel dataset of 314 annotated seminal publications and literature reviews along with their metadata (including DOIs, titles, authors, and abstracts), which we call TrueImpactDataset. We described how this dataset was built, provided a detailed analysis of the dataset and discussed the properties an ideal dataset for validating research evaluation metrics should have. We share this dataset with the research community[7] and hope it will be useful to others and will perhaps inspire creating a true ground truth evaluation set.

References

Giovanni Abramo, Ciriaco Andrea DAngelo, and Flavia Di Costa. Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? *Scientometrics*, 84(3):821–833, 2010.

Nancy J Adler and Anne-Wil Harzing. When knowledge wins: Transcending the sense and nonsense of academic rankings. *Academy of Management Learning & Education*, 8(1):72–95, 2009.

Robert Adler, John Ewing, and Peter Taylor. Citation statistics. *Statistical Science*, 24(1):1, 2009.

D W Aksnes. Characteristics of highly cited papers. *Research Evaluation*, 12 (3):159–170, 2003. ISSN 09582029. doi: 10.3152/147154403781776645.

Benjamin M. Althouse, Jevin D. West, Carl T. Bergstrom, and Theodore Bergstrom. Differences in Impact Factor Across Fields and Over Time. *Journal of the American Society for Information Science and Technology*, 60(1):27–34, 2009. ISSN 14923831. doi: 10.1002/asi.20936.

John Antonakis, Nicolas Bastardoz, Yonghong Liu, and Chester A Schriesheim. What makes articles highly cited? *The Leadership Quarterly*, 25(1):152–179, 2014.

Australian Research Council. Excellence in research for australia: Era 2015 evaluation handbook. Technical report, 2015.

Marc Bertin, Iana Atanassova, Yves Gingras, and Vincent Larivière. The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, 67(1):164–177, 2016.

Lutz Bornmann. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of informetrics*, 8(4):895–903, 2014.

[7] http://trueimpactdataset.semantometrics.org

Lutz Bornmann. Usefulness of altmetrics for measuring the broader impact of research: A case study using data from plos and f1000prime. *Aslib Journal of Information Management*, 67(3):305–319, 2015.

Lutz Bornmann and Hans-Dieter Daniel. Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3):391–392, 2005.

Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1): 45–80, 2008.

Lutz Bornmann and Robin Haunschild. Which people use which scientific papers? an evaluation of data from f1000 and mendeley. *Journal of informetrics*, 9(3):477–487, 2015.

Lutz Bornmann and Robin Haunschild. Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? *Scientometrics*, 110(2):937–943, 2017.

Lutz Bornmann and Loet Leydesdorff. Does quality and content matter for citedness? a comparison with para-textual factors and over time. *Journal of Informetrics*, 9(3):419–429, 2015.

Lutz Bornmann, Irina Nast, and Hans-Dieter Daniel. Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? a quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication. *Scientometrics*, 77 (3):415–432, 2008.

Linda Butler. Using a balanced approach to bibliometrics: quantitative performance measures in the australian research quality framework. *Ethics in Science and Environmental Politics*, 8(1):83–92, 2008.

Ciriaco Andrea D'Angelo and Giovanni Abramo. Publication Rates in 192 Research Fields of the Hard Sciences. In *Proceedings of the 15th ISSI Conference*, pages 915–925, 2015.

Olivier Francois. Arbitrariness of peer review: A bayesian analysis of the nips experiment. *arXiv preprint arXiv:1507.06411*, 2015.

Eugene Garfield. The meaning of the impact factor. *International Journal of Clinical and Health Psychology*, 3(2), 2003.

Nigel Harwood. An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, 41(3):497–518, 2009.

Anne-Wil Harzing. Microsoft Academic (Search): a Phoenix arisen from the ashes? page 11, 2016.

Anne-Wil Harzing and Satu Alakangas. Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016a. doi: 10.1007/s11192-015-1798-9.

Anne-Wil Harzing and Satu Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016b.

Stefanie Haustein and Vincent Larivière. Mendeley as a source of readership by students and postdocs? evaluating article usage by academic status. *Proceedings of the IATUL Conferences*, 2014.

Stefanie Haustein, Isabella Peters, Judit Bar-Ilan, Jason Priem, Hadas Shema, and Jens Terliesner. Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, 101(2):1145–1163, 2014.

Zhigang Hu, Chaomei Chen, and Zeyuan Liu. The recurrence of citations within a scientific article. In *ISSI*, 2015.

Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. Peer review in scientific publications: benefits, critiques, & a survival guide. *Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine*, 25(3):227, 2014.

Petr Knoth and Drahomira Herrmannova. Towards semantometrics: A new semantic similarity based measure for assessing a research publication's contribution. *D-Lib Magazine*, 20(11):8, 2014.

Gabriel Kreiman and John HR Maunsell. Nine criteria for a measure of scientific output. *Frontiers in computational neuroscience*, 5(48):11, 2011.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Michael H MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61(1):1–12, 2010a.

Michael H MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61(1):1–12, 2010b.

Nabeil Maflahi and Mike Thelwall. When are readership counts as useful as citation counts? scopus versus mendeley for lis journals. *Journal of the Association for Information Science and Technology*, 67(1):191–199, 2016. DOI: 10.1002/asi.23369.

Marie E McVeigh and Stephen J Mann. The journal impact factor denominator: defining citable (counted) items. *Jama*, 302(10):1107–1109, 2009.

Lokman I Meho. The rise and rise of citation analysis. *Physics World*, 20(1): 32, 2007.

Ehsan Mohammadi, Mike Thelwall, Stefanie Haustein, and Vincent Larivière. Who reads research articles? an altmetrics analysis of mendeley user categories. *Journal of the Association for Information Science and Technology*, 66(9):1832–1846, 2015.

Ehsan Mohammadi, Mike Thelwall, Kayvan Kousha, et al. Can mendeley bookmarks reflect readership? a survey of user motivations. *JASIST*, 67(5): 1198–1209, 2016.

Olgica Nedić and Aleksandar Dekanski. Priority criteria in peer review of scientific articles. *Scientometrics*, 107(1):15–26, 2016.

Natsuo Onodera and Fuyuki Yoshikane. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4):739–764, 2015.

Robert M. Patton, Christopher G. Stahl, and Jack C. Wells. Measuring Scientific Impact Beyond Citation Counts. *D-Lib Magazine*, 22(9/10):5, 2016.

Heather Piwowar and Jason Priem. The power of altmetrics on a cv. *Bulletin of the American Society for Information Science and Technology*, 39(4):10–13, 2013. DOI: 10.1002/bult.2013.1720390405.

David Pride and Petr Knoth. Incidental or influential? - challenges in automatically detecting citation importance using publication full texts. In *Theory and Practice of Digital Libraries (TPDL) 2017*, Thessaloniki, Greece, 2017.

Jason Priem. Altmetrics. In Blaise Cronin and Cassidy R Sugimoto, editors, *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*, chapter 14, pages 263–288. MIT Press, Cambridge, MA, 2014.

Jason Priem, Heather A Piwowar, and Bradley M Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745*, 2012.

REF 2014. Panel criteria and working methods. Technical Report January 2012, 2012.

Research Excellence Framework. Panel criteria and working methods. Technical report, 2012.

Research Excellence Framework. Research Excellence Framework (REF) 2014 Units of Assessment. `http://www.ref.ac.uk/panels/unitsofassessment/`, 2014. Accessed: 2016-11-11.

Martin Ricker. Letter to the editor: About the quality and impact of scientific articles. *Scientometrics*, 111(3):1851–1855, 2017.

Per Ottar Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(February):498–502, 1997.

Robert J Sternberg and Tamara Gordeeva. The anatomy of impact: What makes an article influential? *Psychological Science*, 7(2):69–75, 1996.

Jaime A Teixeira da Silva and Judit Dobránszki. Problems with traditional science publishing and finding a wider niche for post-publication peer review. *Accountability in research*, 22(1):22–40, 2015.

Tertiary Education Commission. Performance-based research fund: Quality evaluation guidelines 2012. Technical report, 2013.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110. Association for Computational Linguistics, 2006.

Mike Thelwall and Kayvan Kousha. Web indicators for research evaluation. part 1: Citations and links to academic articles from the web. *El profesional de la información*, 24(5):587–606, 2015a.

Mike Thelwall and Kayvan Kousha. Web indicators for research evaluation. part 2: Social media metrics. *El profesional de la información*, 24(5):607–620, 2015b.

Mike Thelwall and Pardeep Sud. Mendeley readership counts: An investigation of temporal and disciplinary differences. *Journal of the Association for Information Science and Technology*, 67(12):3036–3050, 2016.

Thomson Reuters. Journal citation reports – journal source data. `http://admin-apps.webofknowledge.com/JCR/help/h_sourcedata.`

`htm#sourcedata`. Version: 2012-05-22, Accessed: 2017-01-26.

Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Richard Van Noorden, Brendan Maher, and Regina Nuzzo. The top 100 papers. *Nature*, 514(7524):550, 2014.

Ludo Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391, 2016.

Xiaojun Wan and Fang Liu. Are all literature citations equally important? automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, 65(9):1929–1938, 2014.

Ryan Whalen, Yun Huang, Anup Sawant, Brian Uzzi, and Noshir Contractor. Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science. *ASCW'15 Workshop at Web Science 2015*, pages 6–8, 2015.

James Wilsdon, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.* 2015. ISBN 1902369273. doi: 10.13140/RG.2.1.4929.1363.

Rui Yan, Congrui Huang, Jie Tang, Yan Zhang, and Xiaoming Li. To Better Stand on the Shoulder of Giants. In *Proceedings of the 12th Joint Conference on Digital Libraries*, pages 51–60, Washington, DC, 2012. ACM. ISBN 9781450311540.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427, 2015.

A Experiment results

Discipline	p **(citations)**	p **(readership)**	**Total**
Geography, Environmental Studies and Archaeology	0.3404	0.2081	8
Biological Sciences	0.1748	0.4956	17
Computer Science and Informatics	0.0895	0.4517	43
Mathematical Sciences	0.2549	0.2518	14
Earth Systems and Environmental Sciences	0.1162	0.1645	18
Business and Management Studies	0.1191	0.1577	19
Physics	0.3819	0.1679	26
Education	0.1162	0.2146	26
Psychology, Psychiatry and Neuroscience	0.2443	0.2293	9
Politics and International Studies	0.2007	0.4275	6
Electrical and Electronic Engineering, Metallurgy and Materials	0.4260	0.3397	16
Sociology	0.4302	0.3955	7
Classics	0.1265	0.2113	4
Art and Design: History, Practice and Theory	0.2702	0.4565	5
Social Work and Social Policy	0.0910	0.3365	6
Economics and Econometrics	0.1525	0.3977	8
General Engineering	0.2079	0.1453	4
Anthropology and Development Studies	0.2920	0.2850	4
Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.2439	0.2015	4
Modern Languages and Linguistics	0.1557	0.1154	4
Public Health, Health Services and Primary Care	0.2056	0.1906	6
Total	-	-	254

Table 13 Results of independent one-tailed t-test performed using citation and readership counts on all disciplines separately.

Discipline	**Acc.**	**Opt.**	**Base.**	**Opt.** t	**TN**	**TP**	**FN**	**FP**	**Total**
Geography, Environmental Studies and Archaeology	0.3750	0.7500	0.5000	41	2	1	3	2	8
Biological Sciences	0.2941	0.6471	0.5294	50	4	1	7	5	17
Computer Science and Informatics	0.3023	0.6279	0.5349	50	7	6	17	13	43
Mathematical Sciences	0.5714	0.6429	0.5714	14	1	7	1	5	14
Earth Systems and Environmental Sciences	0.3333	0.6667	0.5000	59	3	3	6	6	18
Business and Management Studies	0.4737	0.6842	0.5263	197	6	3	6	4	19
Physics	0.6154	0.6154	0.5000	916	12	4	9	1	26
Education	0.3846	0.6923	0.5769	19	3	7	8	8	26
Psychology, Psychiatry and Neuroscience	0.4444	0.6667	0.5556	31	1	3	2	3	9
Politics and International Studies	0.6667	0.6667	0.5000	389	3	1	2	0	6
Electrical and Electronic Engineering, Metallurgy and Materials	0.6250	0.6875	0.5000	50	5	5	3	3	16
Sociology	0.7143	0.8571	0.5714	2	2	3	1	1	7
Classics	0.7500	1.0000	0.5000	25	2	1	1	0	4
Art and Design: History, Practice and Theory	0.2000	0.6000	0.6000	0	0	1	2	2	5
Social Work and Social Policy	0.5000	0.8333	0.5000	17	2	1	2	1	6
Economics and Econometrics	0.6250	0.7500	0.5000	119	3	2	2	1	8
General Engineering	0.5000	0.7500	0.5000	69	1	1	1	1	4
Anthropology and Development Studies	0.0000	0.5000	0.5000	0	0	0	2	2	4
Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.7500	0.7500	0.5000	2138	2	1	1	0	4
Modern Languages and Linguistics	0.7500	1.0000	0.5000	38	2	1	1	0	4
Public Health, Health Services and Primary Care	0.3333	0.6667	0.5000	2	1	1	2	2	6
All	0.4528	0.6811	-	-	62	53	79	60	254

Table 14 Classification results using citation counts as a feature, performed on all disciplines separately. The columns TN, TP, FN and FP show the number of true negatives (papers correctly predicted as review), true positives (papers correctly predicted as seminal), false negatives (seminal papers incorrectly predicted as review) and false positives (review papers incorrectly predicted as seminal), respectively. The column "Opt." shows accuracy achieved with the optimal model and column "Base." shows accuracy of the baseline model.

Discipline	**Acc.**	**Opt.**	**Base.**	**Opt.** t	**TN**	**TP**	**FN**	**FP**	**Total**
Geography, Environmental Studies and Archaeology	0.0000	0.5000	0.5000	0	0	0	4	4	8
Biological Sciences	0.4118	0.5882	0.5294	123	6	1	7	3	17
Computer Science and Informatics	0.3953	0.5349	0.5349	0	0	17	6	20	43
Mathematical Sciences	0.0714	0.5714	0.5714	0	0	1	7	6	14
Earth Systems and Environmental Sciences	0.7778	0.7778	0.5000	96	5	9	0	4	18
Business and Management Studies	0.6316	0.6316	0.5263	256	7	5	4	3	19
Physics	0.2308	0.6154	0.5000	4	4	2	11	9	26
Education	0.6154	0.6154	0.5769	1	4	12	3	7	26
Psychology, Psychiatry and Neuroscience	0.3333	0.6667	0.5556	21	1	2	3	3	9
Politics and International Studies	0.3333	0.6667	0.5000	1	1	1	2	2	6
Electrical and Electronic Engineering, Metallurgy and Materials	0.5000	0.6250	0.5000	43	7	1	7	1	16
Sociology	0.4286	0.7143	0.5714	40	1	2	2	2	7
Classics	0.7500	0.7500	0.5000	1	2	1	1	0	4
Art and Design: History, Practice and Theory	0.2000	0.6000	0.6000	0	0	1	2	2	5
Social Work and Social Policy	0.1667	0.5000	0.5000	0	0	1	2	3	6
Economics and Econometrics	0.5000	0.6250	0.5000	77	3	1	3	1	8
General Engineering	0.5000	1.0000	0.5000	82	1	1	1	1	4
Anthropology and Development Studies	0.7500	0.7500	0.5000	15	1	2	0	1	4
Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.0000	0.5000	0.5000	0	0	0	2	2	4
Modern Languages and Linguistics	0.5000	1.0000	0.5000	36	1	1	1	1	4
Public Health, Health Services and Primary Care	0.3333	0.6667	0.5000	8	0	2	1	3	6
All	0.4213	0.6260	-	-	44	63	69	78	254

Table 15 Classification results using reader counts as a feature, performed on all disciplines separately. The columns TN, TP, FN and FP show the number of true negatives (papers correctly predicted as review), true positives (papers correctly predicted as seminal), false negatives (seminal papers incorrectly predicted as review) and false positives (review papers incorrectly predicted as seminal), respectively. The column "Opt." shows accuracy achieved with the optimal model and column "Base." shows accuracy of the baseline model.

Year	p **(citations)**	p **(readership)**	**Total**
1999	0.3738	0.1951	8
2000	0.1706	0.0555	10
2001	0.1988	0.3102	15
2003	0.1096	0.3459	9
2004	0.4157	0.1629	10
2005	0.2115	0.3178	17
2006	0.3230	0.2259	14
2007	0.1570	0.1482	15
2008	0.2112	0.4029	14
2009	0.1199	0.0531	11
2010	0.1098	0.3501	21
2011	0.2064	0.2207	18
2012	0.1154	0.4622	17
2013	0.4370	0.1918	19
2014	0.2785	0.0731	13
2015	0.4661	0.1684	11
2016	0.0842	0.3098	17
Total	-	-	239

Table 16 Results of independent one-tailed t-test performed using citation and readership counts on all publication years separately.

Year	**Acc.**	**Opt.**	**Base.**	**Opt.** t	**TN**	**TP**	**FN**	**FP**	**Total**
1999	0.7500	0.7500	0.7500	0	0	6	0	2	8
2000	0.6000	0.7000	0.7000	0	0	6	1	3	10
2001	0.1333	0.6000	0.5333	3	1	1	7	6	15
2003	0.6667	0.8889	0.5556	374	3	3	2	1	9
2004	0.3000	0.7000	0.5000	35	2	1	4	3	10
2005	0.4706	0.5882	0.5882	472	8	0	7	2	17
2006	0.5714	0.5714	0.5714	1559	7	1	5	1	14
2007	0.6667	0.6667	0.6000	37	5	5	1	4	15
2008	0.4286	0.7143	0.5000	197	2	4	3	5	14
2009	0.4545	0.5455	0.6364	214	5	0	4	2	11
2010	0.6190	0.7143	0.5714	1105	11	2	7	1	21
2011	0.5000	0.6667	0.5556	59	3	6	4	5	18
2012	0.7059	0.7059	0.6471	633	11	1	5	0	17
2013	0.6316	0.7895	0.7895	240	12	0	4	3	19
2014	0.6923	0.6923	0.7692	64	9	0	3	1	13
2015	0.6364	0.7273	0.7273	96	7	0	3	1	11
2016	0.5882	0.7059	0.5882	2	9	1	6	1	17
All	0.5523	0.6862	-	-	95	37	66	41	239

Table 17 Classification results using citation counts as a feature, performed on all years separately. The columns TN, TP, FN and FP show the number of true negatives (papers correctly predicted as reivew), true positives (papers correctly predicted as seminal), false negatives (seminal papers incorrectly predicted as review) and false positives (review papers incorrectly predicted as seminal), respectively. The column “Opt.” shows accuracy achieved with the optimal model and column “Base.” shows accuracy of the baseline model.

Year	**Acc.**	**Opt.**	**Base.**	**Opt. t**	**TN**	**TP**	**FN**	**FP**	**Total**
1999	0.5000	0.7500	0.7500	0	0	4	2	2	8
2000	0.6000	0.7000	0.7000	0	0	6	1	3	10
2001	0.5333	0.6667	0.5333	57	3	5	3	4	15
2003	0.2222	0.5556	0.5556	0	0	2	3	4	9
2004	0.6000	0.6000	0.5000	15	3	3	2	2	10
2005	0.6471	0.6471	0.5882	327	9	2	5	1	17
2006	0.2143	0.5714	0.5714	39	3	0	6	5	14
2007	0.2000	0.6000	0.6000	10	3	0	6	6	15
2008	0.5000	0.5714	0.5000	2775	6	1	6	1	14
2009	0.4545	0.5455	0.6364	382	5	0	4	2	11
2010	0.5714	0.6190	0.5714	326	11	1	8	1	21
2011	0.3889	0.6111	0.5556	1	2	5	5	6	18
2012	0.4118	0.6471	0.6471	41	7	0	6	4	17
2013	0.7895	0.8421	0.7895	823	14	1	3	1	19
2014	0.6154	0.6923	0.7692	123	8	0	3	2	13
2015	0.7273	0.8182	0.7273	1028	7	1	2	1	11
2016	0.5882	0.6471	0.5882	35	9	1	6	1	17
All	0.5105	0.6527	-	-	90	32	71	46	239

Table 18 Classification results using reader counts as a feature, performed on all years separately. The columns TN, TP, FN and FP show the number of true negatives (papers correctly predicted as review), true positives (papers correctly predicted as seminal), false negatives (seminal papers incorrectly predicted as review) and false positives (review papers incorrectly predicted as seminal), respectively. The column "Opt." shows accuracy achieved with the optimal model and column "Base." shows accuracy of the baseline model.