Emerald Insight

Aslib Journal of Information Management

Article information:

Users who downloaded this article also downloaded:

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

Usefulness of altmetrics for measuring the broader impact of research
A case study using data from PLOS and F1000Prime

Lutz Bornmann

Division for Science and Innovation Studies,
Administrative Headquarters of the Max Planck Society, Munich, Germany

Abstract

Purpose – The purpose of this case study is to investigate the usefulness of altmetrics for measuring the broader impact of research.

Design/methodology/approach – This case study is based on a sample of 1,082 the Public Library of Science (PLOS) journal articles recommended in F1000. The data set includes altmetrics which were provided by PLOS. The F1000 data set contains tags on papers which were assigned by experts to characterise them.

Findings – The most relevant tag for altmetric research is "good for teaching", as it is assigned to papers which could be of interest to a wider circle of readers than the peers in a specialised area. One could expect papers with this tag to be mentioned more often on Facebook and Twitter than those without this tag. The results from regression models were able to confirm these expectations: papers with this tag show significantly higher Facebook and Twitter counts than papers without this tag. This clear association could not be seen with Mendeley or Figshare counts (that is with counts from platforms which are chiefly of interest in a scientific context).

Originality/value – The results of the current study indicate that Facebook and Twitter, but not Figshare or Mendeley, might provide an indication of which papers are of interest to a broader circle of readers (and not only for the peers in a specialist area), and could therefore be useful for the measurement of the societal impact of research.

Keywords Facebook, Altmetrics, Mendeley, Twitter, F1000, Figshare

Paper type Research paper

1. Introduction

While scientometricians have chiefly used citation scores as an instrument for the measurement of the impact of research, in recent years altmetrics (short for alternative metrics) have provided a further attractive possibility for measuring the impact of research. "Altmetrics refers to data sources, tools, and metrics (other than citations) that provide potentially relevant information on the impact of scientific outputs (e.g. the number of times a publication has been tweeted, shared on Facebook, or read in Mendeley). Altmetrics opens the door to a broader interpretation of the concept of impact and to more diverse forms of impact analysis" (Waltman and Costas, 2014, p. 433). Whereas citations allow a measurement of the impact of research on research itself, one assumption is that altmetrics can also capture the impact beyond the realm of science. An overview of the various available altmetrics can be found in

Claussen *et al.* (2013). For example, Twitter (www.twitter.com) is a popular microblogging application, whose data can be used as an altmetric. This application allows the user to post short messages (tweets) of up to 140 characters. Tweets can be sent to other users and linked to web sites or scientific papers (Darling *et al.*, 2013). Other sources of altmetrics are Mendeley, Facebook, and Figshare.

There are already a number of studies concerning altmetrics in scientometrics. An overview of these studies can be found, for example, in Bar-Ilan *et al.* (2014), Bornmann (2014a), Haustein (2014), and Priem (2014). Many researchers in scientometrics (and beyond) are asking for the meaning of altmetrics: for Zahedi *et al.* (2014) altmetrics is still in its infancy and "at the moment, we don't yet have a clear definition of the possible meanings of altmetric scores". According to Taylor and Plume (2014) "altmetrics hold great promise as a source of data, indicators and insights about online attention, usage and impact of published research outputs. What is currently less certain is the underlying nature of what is being measured by current indicators" (p. 19). In order to investigate their underlying nature, most of the studies have investigated the relationship between altmetrics and citation counts (Bornmann, in press-a). The computed correlation coefficients are expected to provide an indication on what is actually measured by altmetrics. However, coefficients for the correlation of altmetrics and citation counts cannot sufficiently answer this question. An altmetric with high correlation could be seen as "not random" (Thelwall, 2014), but equally as a "not so" alternative metric to citation counts. A low correlation could mean that altmetrics measure other dimensions of impact than traditional metrics (Dinsmore *et al.*, 2014). Thus, for a satisfying answer to the question of what is actually measured more suitable data are necessary.

In January 2002, F1000 was launched as a new type of post-publication peer-review system, in which around 5,000 experts, so-called "Faculty members", were asked "to identify, evaluate and comment on the most interesting papers they read for themselves each month – regardless of the journal in which they appear" (Wets *et al.*, 2003, p. 251). The Faculty members also attached tags to the papers indicating their relevance for science, as well as other purposes. Because a tag characterises a paper, the relation between differently tagged papers and altmetric scores can be investigated. For example, a tag of particular interest for altmetrics research is "good for teaching". Papers can be marked by Faculty members in this way if they represent a key paper in a field, are well written, provide a good overview of a topic, and/or are well suited as literature for students. Thus, one can expect that papers tagged as "good for teaching" can not only be understood by people studying that topic but also by people who are not researchers (within a specific field). In this study, the "good for teaching" tag is seen as a marker of societal impact, and this assumption is investigated with altmetric scores.

The present study is based on a data set with papers (and their assessments and tags from Faculty members) extracted from F1000 and limited to papers published by the PLOS (Public Library of Science, a non-profit publisher) and associated PLOS article level metrics (ALMs) (including altmetric data). ALMs at PLOS do not only contain altmetrics, but also other metrics, such as citations from Scopus and downloads. Using the F1000 tags and PLOS ALMs, this case study explores the question whether altmetrics are able to measure some kind of societal impact.

The results of this case study are compared with those of Bornmann (2014c) who used the same F1000 data set matched with an altmetrics data set from Altmetric (www.altmetric.com/) – a start-up that focuses on making ALMs available. Whereas the PLOS data set offers a wide range of ALMs, Bornmann (2014c) included only two

altmetrics: the total Altmetric score and Twitter counts. That means both studies investigate the same questions based on the F1000 data set, but used altmetrics data from different sources.

2. Peer ratings provided by F1000

F1000 is a post-publication peer review system of articles from biological and medical journals. F1000 is part of the Science Navigation Group. This group of independent companies publishes and develops information services for the biomedical community. F1000 started with F1000 Biology, which was launched in 2002. In 2006, F1000 Medicine was added. Both services were merged in 2009 and constitute the current F1000 database. Papers for F1000 are selected by a global "Faculty" of leading scientists and clinicians who are nominated by peers. The Faculty currently numbers more than 5,000 experts worldwide, assisted by further associates, who are organised into more than 40 subjects. Faculty members can choose and evaluate any article that interests them. They rate the articles and explain their importance (F1000, 2012). Only a restricted set of articles from the medical and biological journals covered is reviewed (Wouters and Costas, 2012; Kreiman and Maunsell, 2011). Although many papers published in popular and high-profile journals (e.g. *Nature*, *New England Journal of Medicine*, *Science*) are evaluated, 85 per cent of the papers selected come from specialised or less well-known journals (Wouters and Costas, 2012).

The papers selected for F1000 are rated by the Faculty members as "Good", "Very good", or "Exceptional" which is equivalent to scores of 1, 2, or 3, respectively. In many cases a paper is assessed not just by one member but by several. According to the results of Bornmann (in press-b) "the papers received between one and 20 recommendations from different Faculty members. Most of the papers (around 94 per cent) have one recommendation (around 81 per cent) or two recommendations (around 13 per cent) by Faculty members". The FFa (F1000 Article Factor), given as a total score in the F1000 database, is the sum of the scores from the different recommendations for a publication. Besides making recommendations, Faculty members also tag publications with classifications, as for example (see http://f1000. com/prime/about/whatis/how):

- Clinical trial (non-RCT): investigates the effects of an intervention (but neither randomised nor controlled) in human subjects.

- Confirmation: the findings of the article validate previously published data or hypotheses.

- Controversial: findings of the article either challenge the established dogma in a given field, or require further validation before they can be considered irrefutable.

- Good for teaching: a key article in that field and/or a particularly well-written article that provides a good overview of a topic or is an excellent example of which students should be aware. The "good for teaching" tag is relatively new for F1000Prime; it was introduced only in 2011.

- Interesting hypothesis: proposes a novel model or hypothesis that the recommending Faculty member found worthy of comment.

- New finding: presents original data, models, or hypotheses.

- Novel drug target: the article suggests a specific, new therapeutic target for drug discovery (rather than a new drug).

- Refutation: the findings of the article disprove previously published data or hypotheses (where a specific finding is being refuted, a reference citation is normally required).

- Technical advance: introduces a new practical/theoretical technique, or novel use or combination of an existing technique or techniques.

The classifications of the papers are intended to serve as additional filters rather than part of the rating by a Faculty member. The tags are very useful because they are assigned by an expert. Whereas literature databases (e.g. Web of Science, Thomson Reuters) cannot be searched for negative results, or clinical practice changing papers, the human expert-assigned tags enable this in F1000Prime.

Overall, the F1000 database is regarded as an aid for scientists to receive pointers to the most relevant papers in biomedicine, but also as an important tool for research evaluation purposes. According to Wouters and Costas (2012) "the data and indicators provided by F1000 are without doubt rich and valuable, and the tool has a strong potential for research evaluation, being in fact a good complement to alternative metrics for research assessments at different levels (papers, individuals, journals, etc.)" (p. 14). Thus, it is no wonder that several empirical studies have been published which are based on the F1000 database (Du *et al.*, 2014; Li and Thelwall, 2012; Mohammadi and Thelwall, 2013; Waltman and Costas, 2014; Wardle, 2010).

3. Methods
3.1 Construction of the data set
In January 2014, F1000 provided the author with data on all recommendations (and classifications) and the bibliographic information for the corresponding papers in their system ($n = 149{,}227$ records). The data set is based on single recommendations for papers and not on single papers. Thus, many papers appear multiply in the data set. By using the DOI, the F1000 data set was matched with another data set downloaded from the PLOS homepage (10.6084/m9.figshare.1189396). The PLOS data set (alm_report_2014-03-10.csv) includes all papers ($n = 114{,}093$ records) published by PLOS and a large number of ALMs (see http://article-level-metrics.plos.org/alm-info/). An overview of the numerous ALMs provided by PLOS can be found in Fenner (2013). Matching of the two data sets with the DOI reduced the data set to 3,928 records. The marked reduction in the records with the matching of the data sets can be explained by the high selectivity of F1000 Faculty members in reading and assessing of papers. Furthermore, the PLOS database contains only a small subset of the journals which the members generally read and assess. The 3,928 records relate to the publication years between 2003 and 2013. Since the "Good for teaching" tag, which Faculty members can assign, was not used by F1000 before 2011, the data set for this case study is further reduced to 1,204 records published during the years 2012 and 2013. Since papers which received more than one recommendation from a Faculty member appear multiply in the data set, the 1,204 records refer to 1,082 papers.

In Table I ALMs from PLOS are listed for all papers included in the study. The data set with the PLOS ALMs is published at regular intervals, while the provided ALMs are changed. Table I therefore relates to the core set of metrics which have generally been listed by PLOS for a longer time (with some exceptions, like DataCite). For each ALM, the table gives the proportion of papers in 2012 and 2013 which have non-zero counts. As the results in the table show, the share varies greatly with the ALMs and ranges from 0 per cent with ScienceSeeker up to 100 per cent with PLOS downloads.

ALM	Proportion of papers in 2012 with non-zero counts ($n = 586$)	Proportion of papers in 2013 with non-zero counts ($n = 496$)
Social network		
Facebook	49.5	51.0
Twitter	49.3	62.3
Usage		
CiteULike	31.1	20.2
Figshare	83.6	85.3
Mendeley	84.8	47.4
Citation		
CrossRef	94.7	64.9
DataCite	0.3	0.4
PubMed Central	100.0	100.0
Scopus	76.1	56.9
Comments, blogs, and media		
Nature.com posts	1.4	0.0
PLOS comments	17.2	13.9
Reddit	0.5	0.0
Research blogging	2.0	1.0
ScienceSeeker	0.0	0.0
Wikipedia	5.5	3.4
WordPress	6.8	6.9
Downloads		
PLOS downloads	100.0	100.0
PubMed downloads	81.2	44.6
Recommendations		
F1000 summary metric	100.0	100.0
Note: $n = 1,082$		

Table I.
Proportion of PLOS
papers, which were
published after 2011
and which have
non-zero counts for
the ALMs used

Under the label "social network" and "usage", Table I provides the group of those altmetrics which belong to well-known altmetrics and are the focus of this case study. Unlike another group of altmetrics listed in the table under "comments, blogs and media", the social network and usage altmetrics are available for the majority of papers (except for CiteULike) and thus are not affected by an inflation of zeros which lead to problems with the statistical analysis (calculation of regression models) (Long and Freese, 2006). Count response models with far more zeros than expected result in incorrect parameter estimates as well as biased standard errors (Hardin and Hilbe, 2012). Of the five social network and usage metrics named in Table I, four are included in the current study: Facebook, Figshare, Mendeley, and Twitter. CiteULike is disregarded in the case study, since it is a platform similar to Mendeley, and Mendeley generally covers the literature more thoroughly than CiteULike (Bornmann, 2014a).

The four altmetrics included in the analysis are outlined in the following. More comprehensive descriptions can be found in the overview works on altmetrics referred to in the introduction. Facebook (facebook.com) is the most often used social media site, where users can create a profile page and interact with other users (Bik and Goldstein, 2013). Since these exchanges with other users may include references to publications, there is the possibility of counting these references and using them as an altmetric.

Figshare (figshare.com) is a relatively new application (since 2011), in which users can store their research output (such as papers, data, tables, and graphics) in a repository and exchange among each other. Many PLOS papers provide supplementary data sets via Figshare. Since Figshare tracks the download statistics for the research output listed, these data are used as a source for altmetrics (Hahnel, 2013).

Online reference managers combine social bookmarking service and reference management functionalities in one platform. They can be seen as the scientific variant of social bookmarking platforms, in which users can save and tag web resources (e.g. blogs or web sites). A popular online reference manager is Mendeley (mendeley.com), which was launched in 2008 (Li *et al.*, 2012). The platform allows users to save and organise literature, share literature with other users, as well as save keywords and comments on a publication (or assign tags to them) (Bar-Ilan *et al.*, 2014; Haustein *et al.*, 2014b). User counts, which provide the number of readers via the saves of publications, serve as altmetrics (Li *et al.*, 2012).

With micro-blogging, users send short messages to other users of a platform. A popular microblogging platform is Twitter (twitter.com), which was founded in 2006. The frequency of the tweets referring to a particular paper can be determined, and used as a source for altmetrics.

3.2 Statistical procedure and software used

The statistical software package Stata 13.1 (www.stata.com/) is used for this case study; in particular, the Stata commands *nbreg*, *margins*, and *coefplot* are used. For the statistical analyses of the data, a series of regression models has been estimated. Since the outcome variables (Facebook, Figshare, Mendeley, and Twitter counts) in the models indicate "how many times something has happened" (Long and Freese, 2006, p. 350), they are count variables. The Poisson distribution is as a rule used to model information on counts. However, this distribution rarely fits in the statistical analysis of altmetric data, due to overdispersion. "That is, the [Poisson] model underfits the amount of dispersion in the outcome" (Long and Freese, 2006, p. 372). The standard model to account for overdispersion is the negative binomial (Hausman *et al.*, 1984). Thus, negative binomial regression models are calculated in the present case study (Hilbe, 2007). The violation of the assumption of independent observations by including several F1000 recommendation scores associated with a paper is considered in the regression models by using the cluster option in Stata (StataCorp., 2013). This option specifies that the scores are independent across papers but are not necessarily independent within the same paper (Hosmer and Lemeshow, 2000, section 8.3). The publication years (2012 and 2013) of the papers were included in the models predicting different counts as exposure time (Long and Freese, 2006). The exposure option provided in Stata takes into account the time that a paper is available for mentions.

In this case study, predicted probabilities are used to make the results easy to understand and interpret. Such predictions are referred to as margins, predictive margins, or adjusted predictions (Bornmann and Williams, 2013; Williams, 2012). The predictions allow an interpretation of the empirical results which goes beyond the statistical significance test. Whereas adjusted predictions can provide a feel for the practical significance of the findings, the regression models illustrate which effects are statistically significant and the direction of the effects.

3.3 The distribution and selection of the tags in the data set and research questions

Table II shows the distribution of the tags across the records in the data set (with multiple occurrences of PLOS papers) or total tag mentions (see "total" line). It is clear

Tag	Absolute numbers	Per cent of tag mentions	Per cent of records
New finding	823	46.05	68.36
Interesting hypothesis	249	13.93	20.68
Confirmation	197	11.02	16.36
Good for teaching	163	9.12	13.54
Technical advance	159	8.90	13.21
Controversial	88	4.92	7.31
Novel drug target	69	3.86	5.73
Refutation	13	0.73	1.08
Systematic review	10	0.56	0.83
Review	8	0.45	0.66
Negative	4	0.22	0.33
Clinical trial (non-RCT)	4	0.22	0.33
Total	1787	100.00	148.42

Notes: $n = 1{,}204$ records, $n = 1{,}787$ tag mentions

Table II.
Tags allocated by
faculty members

that the tags were applied very differently: Whereas, for example, "new finding" comprises about half of the tag mentions, for "controversial" it is only about 5 per cent. In order to make a reliable statement about the validity of the altmetrics, the following statistical analysis does not include all tags, but only those with more than 5 per cent of mentions or allocated to more than 10 per cent of records (see grey marked tags in Table II).

The expectations in the current study in relation to the connection between altmetrics counts and the categorisation of papers with the five selected tags (which are described in further detail in Section 2.1) are as follows: In connection with "new finding", "confirmation" and "interesting hypothesis", it is expected that the Figshare and Mendeley counts for such papers would be higher for those where Faculty member have applied this tag than for those without this tag. These tags chiefly relate to aspects which are relevant in a scientific context, and Figshare and Mendeley are mainly used in this context. With Facebook and Twitter counts, this difference between tagged or untagged papers is not expected since the two platforms do not appeal to a specifically scientific set of users.

We can expect that papers tagged with "good for teaching" would not only be interesting for students, but also for a set of people outside science (or research). These are papers, which are well-written, provide an overview of a topic, and are well suited for teaching. Therefore, higher Facebook and Twitter counts are expected for papers with this tag than for papers without it. With the Mendeley and Figshare counts, this difference is not expected, since these platforms are more strongly oriented to science.

The "technical advance" tag is used for papers that present a new technique or tool (a lab technique/tool or a clinical one) that make an advance on an existing technique. The tag can be used both for research papers and outside, i.e. clinical or fieldwork. Accordingly, no large difference between the four altmetrics is to be expected.

4. Results
In order to examine how Mendeley, Facebook, Twitter, and Figshare counts differ with differently tagged papers, four regression models with four counts as dependent variable and the tags as independent variable were calculated. As Table III shows, each model includes the individual recommendation scores of the Faculty members alongside the tags. This can be used to investigate the influence of the tags on the

Variable	Mean (%)	SD	Minimum	Maximum
Dependent variables				
Mendeley counts (Model 1)	14.87	34.82	0	460
Facebook counts (Model 2)	31.45	341.54	0	11,459
Twitter counts (Model 3)	7.82	32.79	0	800
Figshare counts (Model 4)	7.38	12.69	0	174
Dependent variables				
Tag				
New finding	68%		0	1
Confirmation	16%		0	1
Interesting hypothesis	21%		0	1
Good for teaching	14%		0	1
Technical advance	13%		0	1
Recommendation of Faculty member				
1 "good" (reference category)	51%		0	1
2 "very good"	42%		0	1
3 "exceptional"	7%		0	1
Journal				
PLOS Biology (reference category)	19%		0	1
PLOS Computational Biology	3%		0	1
PLOS Genetics	12%		0	1
PLOS Medicine	3%		0	1
PLOS Neglected Tropical Diseases	2%		0	1
PLOS One	60%		0	1
PLOS Pathogens	1%		0	1
Number of recommendation	$n = 1,204$			
Number of papers	$n = 1,082$			

Table III.
Dependent and independent variables which were included in the four negative binomial regression models

different counts – under the control of the effects of the recommendations. Since the recommendations reflect the quality of the papers, the results of the tags are adjusted for the quality of the papers. In other words: the different results for the tags can hardly be traced back to the differing quality of the papers. The models also include the journals in which the papers were published. This allows the influence of the journal on the counts to be controlled for in the analysis. Thus, for instance, one may expect high-impact journals (e.g. PLOS Biology) to have a larger circle of readers than journals with a lower impact, and that their contributions would also be mentioned more often in social media than the contributions of other journals (Haustein and Siebenlist, 2011).

The results of the regression models are shown in Table IV. The predicted numbers of counts resulting from the models are presented in Figure 1 for the different recommendation scores and in Figure 2 for the different tags. Since the predicted numbers of counts depend on the models with all independent variables, they are calculated for the different tags under control of the recommendation scores (and thus adjusted for quality). As the results of the regression models in Table IV show, the coefficients for the recommendation scores of the Faculty members are statistically significant with the models for the Mendeley and Twitter counts: Papers assessed as "very good" show higher counts than those just rated as "good". The relation between the different recommendation scores and the predicted numbers of counts is presented in Figure 2. As the graphic shows, with Mendeley, Facebook, and Twitter a difference is visible between "good" and "very good", but not between "very good" and

	Model 1: Mendeley	Model 2: Facebook	Model 3: Twitter	Model 4: Figshare	Usefulness of altmetrics
Tag					
New finding	0.05	0.50*	−0.14	0.21	
	(0.36)	(1.98)	(−0.68)	(1.72)	
Confirmation	0.10	0.32	0.33	−0.15	313
	(0.72)	(0.94)	(1.16)	(−1.22)	
Interesting hypothesis	−0.19	0.35	0.08	0.11	
	(−1.48)	(1.57)	(0.45)	(0.87)	
Good for teaching	0.07	0.67**	0.80***	0.23	
	(0.40)	(2.68)	(3.47)	(1.76)	
Technical advance	0.58**	0.10	0.40	0.05	
	(3.20)	(0.42)	(1.52)	(0.31)	
Recommendation of Faculty member					
1 "good" (reference category)					
2 "very good"	0.24*	0.43	0.46*	−0.01	
	(2.23)	(1.79)	(2.39)	(−0.05)	
3 "exceptional"	0.21	0.46	0.32	−0.12	
	(1.01)	(1.28)	(0.98)	(−0.63)	
Journal					
PLOS Biology (reference category)					
PLOS Computational Biology	0.30	−0.39	−0.72*	−0.17	
	(0.45)	(−0.65)	(−2.05)	(−0.64)	
PLOS Genetics	−0.50	−1.37***	−1.07***	−0.19	
	(−1.90)	(−3.83)	(−4.48)	(−0.98)	
PLOS Medicine	−0.95*	0.23	0.81**	−0.46	
	(−1.99)	(0.40)	(2.61)	(−1.38)	
PLOS Neglected Tropical Diseases	−0.98**	3.57***	1.17	−0.22	
	(−2.79)	(3.35)	(1.23)	(−0.74)	
PLOS Pathogens	−0.89***	−0.40	−0.61	−0.09	
	(−3.43)	(−0.75)	(−1.71)	(−0.43)	
PLOS One	−1.15***	−0.36	−0.82***	−0.23	
	(−4.90)	(−1.03)	(−3.54)	(−1.33)	
Constant	−4.35***	−5.02***	−5.51***	−5.61***	Table IV.
	(−17.26)	(−12.68)	(−21.47)	(−26.96)	Results of four negative binomial regression models

Notes: *t* statistics in parentheses; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

"exceptional". The predicted numbers of Figshare counts has almost no relationship – according to the results in Figure 2 – with the recommendation scores.

In the Facebook and Twitter models (Models 2 and 3), the coefficient for "good for teaching" is statistically significant (see Table IV). Correspondingly, Figure 2 shows higher predicted numbers of counts for papers where this tag is set than for those papers where this was not the case. For example, we can expect a paper with this tag to have around nine tweets more than one without – if the paper is rated as "very good" by Faculty members and has no other tags. This result is in agreement with the results of Bornmann (2014c). The result of the present study for "good for teaching" correspond to the expectations and indicate that Facebook and Twitter data can indicate papers which are of interest outside the field of science. Since this effect of the "good for teaching" tag is not demonstrable with Figshare and Mendeley data in this form, such indications are not to be expected with these altmetrics.

Besides the two statistically significant results with the "Good for teaching" tag in the Facebook and Twitter models, the following two statistically significant results appeared in the four regression models (see Table IV): saves by Mendeley users are particularly to be expected when a paper introduces a new practical/theoretical technique (tag: "technical advance"). Since these techniques are an important foundation for work in science, it is understandable that papers with this background would be preferentially saved by Mendeley users. "New finding" is statistically significant in the model with which the Facebook counts are evaluated – although Facebook is mainly a platform for people outside of research. New findings therefore do not lead only to higher traditional citation counts, as Bornmann (2014c) shows, but also to higher Facebook mentions. With reference to Figshare counts it is interesting to note that none of the five tags can be associated with markedly higher or lower altmetric counts (see Figure 2). Downloads from the research output stored on Figshare therefore take place relatively independently of the characteristics of the paper (investigated here).

5. Discussion

Altmetrics are regarded as an attractive possibility for filling the requirement for indicators for broad impact measurement. The data accessibility of altmetrics is already rather good: several organisations (such as ImpactStory, Plum Analytics, and Altmetric) have been founded to collect and provide altmetrics. In addition, altmetrics can also measure the impact of research products other than publications, such as data sets, software, or artwork. As opposed to traditional citations, many altmetrics have the advantage that they permit an impact measurement relatively quickly after the publication of a work. Whereas traditional citations allow a measurement of impact at least three years after the appearance of a publication (Bornmann and Marx, 2014), this is already possible in a (significantly) shorter time scale with many altmetrics.

Since the question, which type of impact can be measured with altmetrics, has not yet been settled, this case study addresses the question with a matched data set of F1000 and PLOS data. The F1000 data set contains tags on papers assigned intellectually by experts to characterise certain characteristics of a paper. "Good for teaching" papers might be of interest not only for students, but also for a wider circle of readers. Particularly on
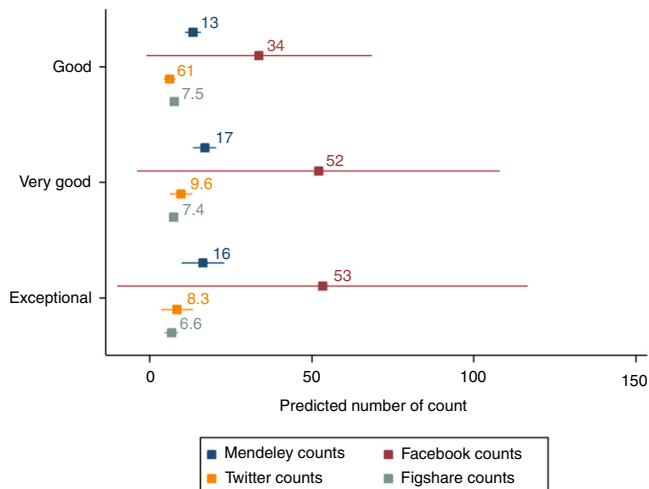

Figure 1.
Predicted numbers of Mendeley, Facebook, Twitter, and Figshare counts with 95 per cent confidence intervals for three individual recommendation scores
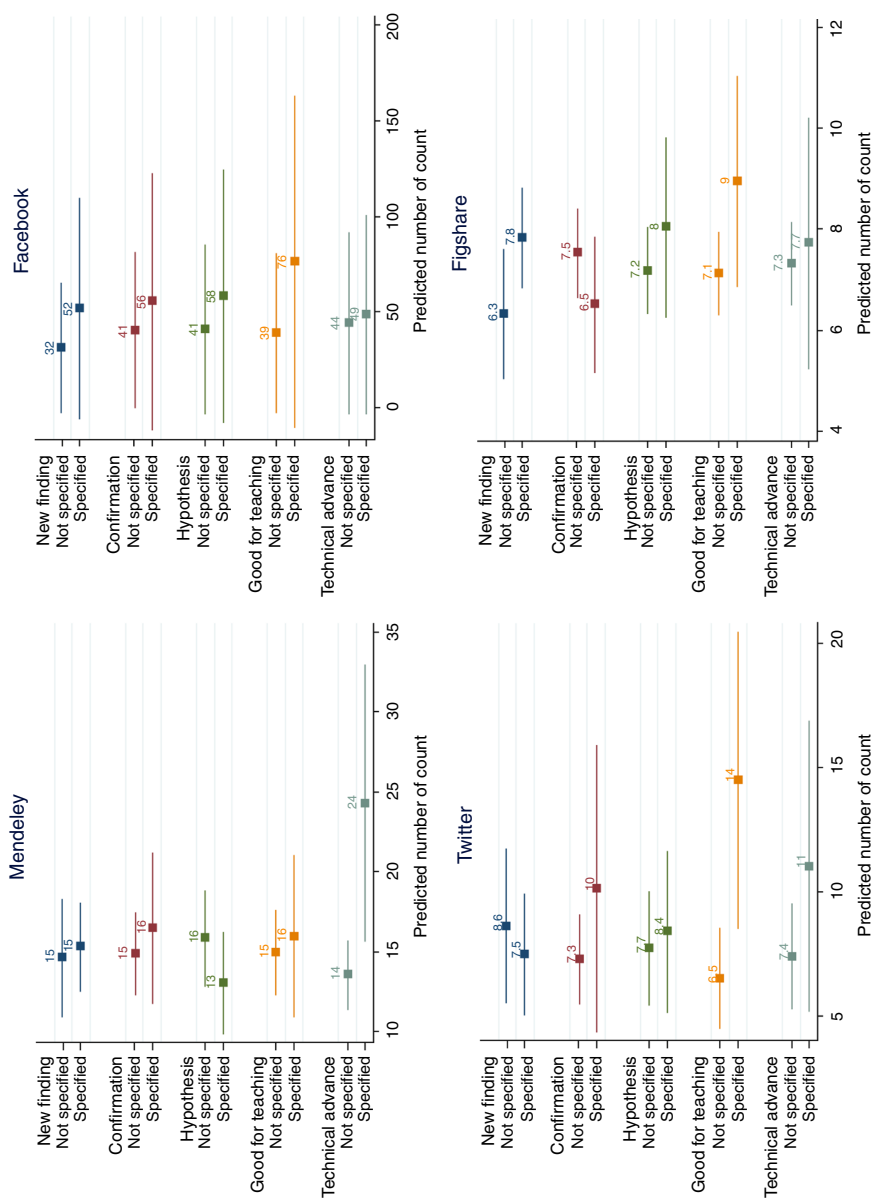
Figure 2.
Predicted numbers
of Mendeley,
Facebook, Twitter,
and Figshare counts
with 95 per cent
confidence intervals
for differently
tagged papers

Facebook and Twitter, one could expect papers with this tag to be mentioned more often than those without this tag. Facebook and Twitter are particularly in use by people outside the area of science: although both platforms belong to the most often used social media, it is generally assumed that only a few scientists are actually professionally active there (Darling *et al.*, 2013; Mahrt *et al.*, 2012; Haustein *et al.*, 2014a).

With respect to the "good for teaching" tag, the results of the current study were able to confirm these expectations: Papers with this tag show significantly higher Facebook and Twitter counts than papers without this tag. This association could not be so clearly seen with Mendeley or Figshare counts (that is with counts from platforms which are chiefly of interest in a scientific context). Especially for Mendeley, one could have expected another result since the majority of Mendeley users are students and students should be particularly interested in papers with this tag. However, the results of this study are in agreement with those of Bornmann (2014c) who investigated also Twitter counts (from Altmetric and not from PLoS) using the F1000 data set. In the results from the Mendeley and Figshare platforms it is interesting to note that papers with the "technical advance" tag have significantly more Mendeley counts than papers without this tag. It seems that interesting new techniques or tools for research which are presented in papers or used for a study can be identified with the help of Mendeley counts.

There are still many research questions on altmetrics to be answered which go well beyond those investigated here. An overview of the numerous further questions may be found in NISO Alternative Assessment Metrics Project (2014), Priem (2014), and Haustein (2014).

Although the present study produced new insights into the broad impact measurement using altmetrics, the study is not without limitations:

(1) The study is based on a very specific subset of papers. The sample is rather limited (only PLOS papers filtered by F1000, $n = 1082$). F1000 is a tool focusing on the biomedical fields and only a very small fraction of all biomedical articles are actually covered. For example, the results show that there are Facebook and Twitter mentions for around 50 per cent (or more) of the papers. These coverage figures are remarkably high for sources that have been reported to have a very low coverage in other – significantly larger – samples (Zahedi *et al.*, 2014; Haustein *et al.*, 2014a; Robinson-García *et al.*, 2014).

(2) Another methodological problem is that the case study is a non-zero analysis, i.e., all papers have F1000 scores. However, we do not know, for example, if there are papers that could be "good for teaching" and are not picked up by F1000. It is not possible to explore whether these "other" papers would have more Twitter activity than the rest.

(3) Papers marked with "good for teaching" can be expected to have an impact beyond science itself, unlike papers without this tag. Although papers marked by this tag have indeed higher Twitter and Facebook counts in this study, it is not clear whether this impact is beyond science. Twitter, Facebook, and other social media are obviously popular outside of science. However, it could not be investigated with the present data set, whether the tweets were sent by users outside the scientific community (e.g. journalists or politicians) or by researchers active on these platforms.

(4) The fourth limitation concerns the Twitter data used in this study: as PLOS started to collect them routinely in June 2012, many tweets from January to May 2012 will not have been captured.

6. Conclusions

Initiated by the wish of governments for a broader impact measurement of scientific activities (Bornmann, 2012, 2013), the field of scientometrics is currently undergoing a process of deep-seated change, which is described by Bornmann (2014b) as a scientific revolution. An impact measurement can no longer involve only traditional citations, but also indicators which can measure the impact of research in other sectors of society (such as the economy, culture, or health services).

The current study makes a contribution to the clarification of the question of the broad impact measured by altmetrics. Even if the results with regard to the use of Twitter and Facebook counts are promising, there are a series of unanswered questions which result from the findings presented here: Why do papers with the "good for teaching" tag receive higher Twitter and Facebook counts than papers without this tag? How does the impact vary with time for papers with this tag: When is the peak for the Facebook and Twitter counts reached (i.e. how soon after publication of a paper can one arrive at a broader impact measurement)? The "controversial" tag could not be included in the present study, because the case number was too low. Controversial papers are relevant mostly for the scientific community (as they belong to the scientific debate). It would be interesting to explore whether they have also a strong visibility in Twitter.

References

Bar-Ilan, J., Shema, H. and Thelwall, M. (2014), "Bibliographic references in web 2.0.", in Cronin, B. and Sugimoto, C.R. (Eds), *Beyond Bibliometrics: Harnessing Multi-Dimensional Indicators of Performance*, MIT Press, Cambridge, MA, pp. 307-325.

Bik, H.M. and Goldstein, M.C. (2013), "An introduction to social media for scientists", *PLoS Biology*, Vol. 11 No. 4, e1001535.

Bornmann, L. (2012), "Measuring the societal impact of research", *EMBO Reports*, Vol. 13 No. 8, pp. 673-676.

Bornmann, L. (2013), "What is societal impact of research and how can it be assessed? A literature survey", *Journal of the American Society of Information Science and Technology*, Vol. 64 No. 2, pp. 217-233.

Bornmann, L. (2014a), "Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics", *Journal of Informetrics*, Vol. 8 No. 4, pp. 895-903.

Bornmann, L. (2014b), "Is there currently a scientific revolution in scientometrics?", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 3, pp. 647-648.

Bornmann, L. (2014c), "Validity of altmetrics data for measuring societal impact: a study using data from Altmetric and F1000Prime", *Journal of Informetrics*, Vol. 8 No. 4, pp. 935-950.

Bornmann, L. (in press-a), "Alternative metrics in scientometrics: a meta-analysis of research into three altmetrics", *Scientometrics*, available at: http://arxiv.org/abs/1407.8010 (accessed 19 March 2015).

Bornmann, L. (in press-b), "Inter-rater reliability and convergent validity of F1000Prime peer review", *Journal of the Association for Information Science and Technology*, available at: http://arxiv.org/abs/1404.0359 (accessed 19 March 2015).

Bornmann, L. and Marx, W. (2014), "How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations", *Scientometrics*, Vol. 98 No. 1, pp. 487-509.

Bornmann, L. and Williams, R. (2013), "How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects", *Journal of Informetrics*, Vol. 7 No. 2, pp. 562-574.

Claussen, J.E., Cooney, P.B., Defilippi, J.M., Fox, S.G., Glaser, S.M., Hawkes, E., Hutt, C., Jones, M. H., Kemp, I.M., Lerner, A., Midway, S.R., Nesbit, S., Osborne-Gowey, J., Roberts, R. and Steward, C. (2013), "Science communication in a digital age: social media and the American fisheries society", *Fisheries*, Vol. 38 No. 8, pp. 359-362.

Darling, E.S., Shiffman, D., Côté, I.M. and Drew, J.A. (2013), "The role of Twitter in the life cycle of a scientific publication", *Ideas in Ecology and Evolution*, Vol. 6 No. 1, pp. 32-43.

Dinsmore, A., Allen, L. and Dolby, K. (2014), "Alternative perspectives on impact: the potential of ALMs and altmetrics to inform funders about research impact", *PLoS Biology*, Vol. 12 No. 11, e1002003.

Du, J., Tang, X. and Wu, Y. (2014), "The effects of research level and article type on the differences between citation metrics and F1000 recommendations", available at: http://arxiv.org/abs/1501.01076 (accessed 16 January 2015).

F1000 (2012), "What is F1000?", available at: http://f1000.com/about/whatis (accessed 25 October 2014).

Fenner, M. (2013), "What can article-level metrics do for you?", *PLoS Biology*, Vol. 11 No. 10, e1001687.

Hahnel, M. (2013), "The reuse factor", *Nature*, Vol. 502 No. 7471, p. 298.

Hardin, J. and Hilbe, J. (2012), *Generalized Linear Models and Extensions*, Stata Press, College Station, TX.

Hausman, J., Hall, B.H. and Griliches, Z. (1984), "Econometric models for count data with an application to the patents R and D relationship", *Econometrica*, Vol. 52 No. 4, pp. 909-938.

Haustein, S. (2014), "Readership metrics", in Cronin, B. and Sugimoto, C.R. (Eds), *Beyond Bibliometrics: Harnessing Multi-Dimensional Indicators of Performance*, MIT Press, Cambridge, MA, pp. 327-344.

Haustein, S. and Siebenlist, T. (2011), "Applying social bookmarking data to evaluate journal usage", *Journal of Informetrics*, Vol. 5 No. 3, pp. 446-457.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D. and Peters, I. (2014a), "Tweets vs Mendeley readers: How do these two social media metrics differ?", *It – Information Technology*, Vol. 56 No. 5, pp. 207-215.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. and Terliesner, J. (2014b), "Coverage and adoption of altmetrics sources in the bibliometric community", *Scientometrics*, Vol. 101 No. 2, pp. 1145-1163.

Hilbe, J.M. (2007), *Negative Binomial Regression*, *Cambridge University Press*, Cambridge.

Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*, John Wiley & Sons, Chichester.

Kreiman, G. and Maunsell, J.H.R. (2011), "Nine criteria for a measure of scientific output", *Frontiers in Computational Neuroscience*, Vol. 5, Article no. 48.

Li, X. and Thelwall, M. (2012), "F1000, Mendeley and traditional bibliometric indicators", *Proceedings of the 17th International Conference on Science and Technology Indicators, Repro-UQAM, Montreal*, pp. 541-551, available at: http://2012.sticonference.org/Proceedings/vol2/Li_F1000_541.pdf (accessed 8 March 2015).

Li, X., Thelwall, M. and Giustini, D. (2012), "Validating online reference managers for scholarly impact measurement", *Scientometrics*, Vol. 91 No. 2, pp. 461-471.

Long, J.S. and Freese, J. (2006), *Regression Models for Categorical Dependent Variables Using Stata*, Stata Press, College Station, TX.

Mahrt, M., Weller, K. and Peters, I. (2012), "Twitter in scholarly communication", in Weller, K., Bruns, A., Burgess, J., Mahrt, M. and Puschmann, C. (Eds), *Twitter and Society*, Peter Lang, New York, NY, pp. 468-483.

Mohammadi, E. and Thelwall, M. (2013), "Assessing non-standard article impact using F1000 labels", *Scientometrics*, Vol. 97 No. 2, pp. 383-395.

NISO Alternative Assessment Metrics Project (2014), "NISO altmetrics standards project white paper", working paper, NISO Alternative Assessment Metrics Project, available at: www. niso.org/apps/group_public/document.php?document_id = 13295&wg_abbrev = altmetrics (accessed 8 July 2014).

Priem, J. (2014), "Altmetrics", in Cronin, B. and Sugimoto, C.R. (Eds), *Beyond Bibliometrics: Harnessing Multi-dimensional Indicators of Performance*, MIT Press, Cambridge, MA, pp. 263-287.

Robinson-García, N., Torres-Salinas, D., Zahedia, Z. and Costas, R. (2014), "New data, new possibilities: exploring the insides of Altmetric.com", *El Profesional de la Información*, Vol. 23 No. 4, pp. 359-366.

Stata Corporation (2013), *Stata Statistical Software: Release 13*, Stata Press, College Station, TX.

Taylor, M. and Plume, A. (2014), "Party papers or policy discussions: an examination of highly shared papers using altmetric data", *Research Trends*, No. 36, pp. 17-20.

Thelwall, M. (2014), "Five recommendations for using alternative metrics in the future UK Research Excellence Framework", The Impact Blog, available at: http://blogs.lse.ac.uk/impactofsocialsciences/2014/10/23/alternative-metrics-future-uk-research-excellence-framework-thelwall/ (accessed 2 January 2015).

Waltman, L. and 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*, Vol. 65 No. 3, pp. 433-445.

Wardle, D.A. (2010), "Do 'Faculty of 1000' (F1000) ratings of ecological publications serve as reasonable predictors of their future impact?", *Ideas in Ecology and Evolution*, Vol. 3, pp. 11-15.

Wets, K., Weedon, D. and Velterop, J. (2003), "Post-publication filtering and evaluation: Faculty of 1000", *Learned Publishing*, Vol. 16 No. 4, pp. 249-258.

Williams, R. (2012), "Using the margins command to estimate and interpret adjusted predictions and marginal effects", *The Stata Journal*, Vol. 12 No. 2, pp. 308-331.

Wouters, P. and Costas, R. (2012), *Users, Narcissism and Control – Tracking the Impact of Scholarly Publications in the 21st Century*, SURFfoundation, Utrecht, available at: www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/Users+narcissism+and+control.pdf (accessed 8 March 2015).

Zahedi, Z., Costas, R. and Wouters, P. (2014), "How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications", *Scientometrics*, Vol. 101 No. 2, pp. 1491-1513.

Corresponding author
Dr Lutz Bornmann can be contacted at: bornmann@gv.mpg.de