Scientometrics (2014) 100:217–225
DOI 10.1007/s11192-013-1215-1

Astrophysics publications on arXiv, Scopus and Mendeley: a case study

Judit Bar-Ilan

Received: 9 September 2013 / Published online: 28 December 2013

Abstract In this study we examined a sample of 100 European astrophysicists and their publications indexed by the citation database Scopus, submitted to the arXiv repository and bookmarked by readers in the reference manager Mendeley. Although it is believed that astrophysicists use arXiv widely and extensively, the results show that on average more items are indexed by Scopus than submitted to arXiv. A considerable proportion of the items indexed by Scopus appear also on Mendeley, but on average the number of readers who bookmarked the item on Mendeley is much lower than the number of citations reported in Scopus. The comparisons between the data sources were done based on the authors and the titles of the publications.

Keywords Altmetrics · Subject-base repositories · Citations · Scopus · arXiv · Mendeley

Introduction

Astrophysicists widely use the e-print server arXiv. arXiv was set up in 1991 as an e-print archive by Paul Ginsparg first for high-energy physics, but it was soon extended to other subfields of physics, including astrophysics. It also covers mathematics and computer science (Grinsparg 1994). Currently arXiv is hosted and funded by the Cornell University Library (arXiv 2013a).

arXiv (www.arxiv.org) is a widely used scholarly repository of eprints: preprints and postprints in several fields and is especially popular in many areas of physics including astrophysics. In 2012, 12,121 items were submitted to arXiv in astrophysics only (arXiv 2013b). As a comparison, Web of Science (WoS) indexed 22,635 items (articles, reviews, proceedings papers and book chapters) published in 2012 under the WoS category

J. Bar-Ilan (✉)
Department of Information Science, Bar-Ilan University, 5290002 Ramat Gan, Israel
e-mail: barilaj@mail.biu.ac.il; Judit.Bar-Ilan@biu.ac.il

Astronomy & Astrophysics. The version of the WoS accessed by us included the Conference Proceedings Citation Index. Larivière et al. (2013) found that about 60 % of all arXiv e-prints are published in a WoS-indexed journal. They considered arXiv submissions in all subject categories for the years 1990–2011. When considering the astrophysics subcategory the percentage of WoS indexed items in arXiv was even higher, around two-thirds, and passed the 70 % mark for journal articles published in 2010. Scopus does not allow for such a fine-grained limitation of subject areas, thus we were not able to estimate the number of astrophysics items indexed by Scopus.

Larivière et al. (2013) also studied the publication and citation patterns of items submitted to the astrophysics section of the arXiv and those indexed by WoS in journals belonging to the NSF category "astronomy and astrophysics". They found that the highest citation rates on WoS (citation data for citations in the publication year and the publication year plus one year) were achieved by items that were also submitted to WoS, but over time the differences between citation rates for these items and items indexed by WoS only diminish.

Mendeley (www.mendeley.com) is a widely used online reference manager. As of August 2013, it contains more than 445 million references of users (a specific item can be referenced by multiple users, and thus counted multiple times). It was founded in 2007 and its first public version was released in 2008 (Mendeley 2013). Mendeley was envisioned as a social reference manager (Henning and Reichelt 2008) that would provide usage-based reputation metrics in addition to being a personal and group reference manager. In Mendeley, users, called "readers" in the system, bookmark items to their personal or group libraries. This information is aggregated, and for each item in the Mendeley database the number of readers who bookmarked the items is publicly available. The readership data can be viewed as an indicator of usage. Usage bibliometrics (Kurtz and Bollen 2010) is a promising complimentary direction to more traditional citation based bibliometrics.

It has been shown in several previous works that there is a significant medium strength correlation between readership counts and citations. Li et al. (2012) showed this for articles published in Science and Nature. They found that the correlation between readership counts and citation counts on WoS and Google Scholar (GS) was 0.559 and 0.592 respectively for articles published in Nature in 2007, and 0.540 and 0.603 for articles published in Science in 2007. Readership and citation counts were collected in July 2010. Bar-Ilan et al. (2012) studied the publication lists of 57 scientometricians and found that Mendeley covered 82 % of the 1,136 publications of these scientometricians found on Scopus and the correlation between Scopus citations and Mendeley readership counts was 0.448. Readership and citation counts were collected in February 2012. Science and Nature articles have high visibility and can be expected to be found on Mendeley, however some of the bibliometricians in our sample were at the beginning of their career, and some of the articles of all the researchers in the sample were not highly cited, still Mendeley had an impressive coverage. This finding was further supported by studies of the Journal of the American Society for Information Science and Technology (Bar-Ilan 2012a, b). These studies show that more than 97 % of the articles published in JASIST in the years 2001 and 2011 were bookmarked by Mendeley readers. The correlations between readership counts and WoS, Scopus and GS citation counts were 0.458, 0.502 and 0.519 respectively. Data were collected in April 2012. In all of the above mentioned studies the correlations were significant. Significance of a finding should not be over emphasized (for a discussion, see Schneider 2013), but in the above-mentioned studies the strength of the correlations seem to be meaningful.

Recently some larger scale studies of Mendeley readership counts versus citation counts were conducted, and these also confirmed significant, medium strength correlations. Mohammadi and Thelwall (in press) studied all social science and humanities publications published in 2008 and indexed by WoS, and collected citation and readership counts in 2012. Correlations were 0.516 for the social sciences and 0.428 for the humanities. Interesting to note, that unlike previous studies, in the study conducted by Mohammadi and Thelwall the median number of Mendeley readership counts was higher than the number of WoS citations. Zahedi et al. (2013) studied a random set of 20,000 WoS publications, and found that Mendeley covered about 37 % of the sample.

There are other altmetric indicators as well (Priem et al. 2010), as shown for example by Priem, Piwowar and Hemminger (Priem et al. 2012), and implemented in services like impactstory (impactstory.org) and altmetric (www.altmetric.com). Wouters and Costas (2012) provide a critical review of some of the available sources and tools. Thelwall et al. (2013) also studied a large set of potential altmetric indicators for a large set of PubMed indexed articles, and found significant associations between citation counts and altmetric counts for a number of these indicators (Twitter, Facebook, wall posts, research highlights, blog mentions, mainstream media and forums). Zahedi et al. (2013) showed that all the potential altmetric indicators except for Mendeley readership counts provided only marginal information.

Li and Thelwall (2012) studied Mendeley reader counts and F1000 (f1000.com) post-publication review assessments with citation counts in Genomics and Genetics. Again nearly all items with high evaluation scores on F1000 were bookmarked by Mendeley readers. Here too the correlations between citation counts and Mendeley readership counts were significant and around 0.68, and were considerably higher than correlations with F1000 evaluations. Formal citations in research blogs in the year of publication were shown to correlate with higher number of future citations (Shema et al. in press). Tweets versus citations were also studied (e.g. Eysenbach 2011; Thelwall et al. 2013).

Research setup

Since the use of arXiv seems to be so prevalent for astrophysicists, we set out to compare the publication lists from arXiv, Scopus and Mendeley for a sample of 100 researchers from EU countries and Israel who published recently in journals indexed by the Web of Science. The sample was selected from a larger sample of more than 500 researchers created for the EU funded ACUMEN project, where only authors for whom we were quite confident that their names were disambiguated properly were included in the subsample. The ACUMEN dataset was built by identifying EU researchers in recent articles published in astrophysics and indexed by the Web of Science. Metadata of the articles in the arXiv repository as of March 2012 were provided to us by Mike Thelwall. Data from Scopus and Mendeley were manually retrieved using the researchers' names. Several versions of the name were used, including firstname–lastname, firstname–middle_inital(s)–lastname and first_initial–middle_initial(s)–lastname. On Mendeley we also searched for the given author using the above variations, but still we could not be sure that all publications were authored by "our" researcher. Since there is no author disambiguation on Mendeley, we only considered publications in the Mendeley database for the given author name that had identical title either with a publication indexed by arXiv or by Scopus. Data from Mendeley were collected between June and August 2012, and data from Scopus were collected during June and July 2012. arXiv and Scopus records were considered identical if the titles matched and the sampled author was one of the authors of the publication. We are aware

that this method might not match all the items, since the arXiv submitted item with almost identical content might not have an identical title with an article published later and indexed by Scopus. On the other hand it is also possible that items with identical titles may differ considerably in content. It should be noted that a similar matching heuristic was employed by Larivière et al. (2013) as well, although we cannot be fully confident that this process did not result in some mistakes. The same is true for matching Mendeley records with Scopus and arXiv records. Larivière et al. (2013) conducted a "macro" study, while the current study can be considered a "micro" study. The advantage of such a micro study is that it is able to highlight extremes and exceptions. Here we not only study citations but readership counts as well.

Results and discussion

The sample researchers were at different stages of their academic career, as can be seen in Table 1. About half of them (49 out of 100) were at an early stage of their career (students, postdocs and assistant professors. The gender distribution was: 18 females and 82 males. Among the women, there was 1 full professor, 3 associate professors, 3 assistant professors and 11 postdocs.

In Table 2 we present the summary of the results per author. We can see that on the average, the number of items indexed by Scopus is larger by almost 50 % than the number of items indexed by arXiv, and only 47 % of the items indexed by Scopus were found in arXiv. On the other hand 30 % of the items submitted to arXiv are not indexed by Scopus, or 70 % of the arXiv submitted items are indexed by Scopus, this is similar to the findings of Larivière et al. (2013). It should be noted that because matching between Scopus and arXiv was based on titles, the actual overlap might be higher. In terms of citations the arXiv submitted items are "responsible" for 54 % of the Scopus citations on average. The overlap between Mendeley with Scopus is smaller than the overlap between arXiv and Mendeley, but the difference is not huge 20.36 items versus 25.04 items. The overlap between the three sources was 14.36 items on the average, which is 27 % of the items indexed by Scopus, and 40 % of the items submitted to arXiv.

We also calculated the sum of readers and citations for the indexed publications of each of the authors. The averages appear in Table 2. One can see that the readership counts are much lower than the citations counts (78.49 vs. 624.09 for arXiv submitted titles, and 90.74 compared with 1168.57 for Scopus indexed items).

Even though there seem to be substantial differences between the sources and between citations and readership counts, the Spearman correlations between them were quite high Table 3.

Table 1 Sample researchers' academic rank distribution

Academic rank	# Researchers in the sample
Full Professor	19
Associate professor/reader/senior lecturer	23
Lecturer, assistant professor	13
Postdoctoral research fellow	30
Student (e.g. PhD or Masters student)	6
Other	9

Table 2 Summary of results per author

	Average	STDEV
arXiv, no. items	36.02	50.36
arXiv-Scopus overlap	25.04	35.28
Sum of citations of arXiv indexed items	624.09	1324.30
arXiv-Mendeley overlap	20.36	33.46
Sum of readers of arXiv indexed items	78.49	138.56
Scopus, no. items	53.92	52.92
Sum of citations of Scopus indexed items	1168.57	2010.60
Scopus-Mendeley overlap	22.02	27.09
Sum of readers of Scopus indexed items	90.74	146.41
Sum of citations of items indexed by Scopus and Mendeley	693.51	1394.10
arXiv-Scopus-Mendeley overlap	14.36	21.68
Sum of readers of items indexed the by the three sources	60.25	109.53
Sum of citations of items indexed the by the three sources	429.40	990.20

Table 3 Correlations between arXiv and Scopus coverage and citation and readership counts

Correlations between:	Spearman's rho	Significance level	95 % confidence interval—bootstrapping 1000 times
# items submitted to arXiv & # items indexed by Scopus	0.763	0.001	(0.661, 0.836)
Scopus citation counts & Mendeley readership counts for items indexed by Scopus	0.786	0.001	(0.688, 0.859)
Scopus citation counts & Mendeley readership counts for items indexed by arXiv, Scopus and Mendeley	0.884	0.001	(0.820, 0.926)

The distributions of arXiv submissions versus Scopus indexed item, Mendeley readership counts versus Scopus citations for Scopus indexed items and Mendeley readership counts versus Scopus citations for arXiv and Scopus indexed items are displayed in Figs. 1, 2 and 3 respectively.

If we calculate article level Spearman correlations between citations and readership counts for items indexed both by Scopus, the correlation is still significant, but quite weak ($r = 0.227$, 95 % confidence intervals using bootstrapping (0.199, 0.253)), which is considerably lower than previous article level results in other fields (Mohammadi and Thelwall in press; Li et al. 2012; Li and Thelwall 2012; Bar-Ilan 2012a, b; Bar-Ilan et al. 2012).

In general we saw that on average the number of items indexed by Scopus was greater than the number of items submitted to arXiv, but if we consider the per-author data we see that for 24 authors there are more arXiv submissions than Scopus records. For seven researchers we did not locate any arXiv submissions, even though their publications were indexed by Scopus, one of them a full professor with 65 items indexed by Scopus. To sum up, it seems that submitting to arXiv is the norm, although not everything is submitted to arXiv.

Moed (2007) studied submissions to the condensed matter subcategory of arXiv, and discussed selection or quality bias in submitting to arXiv. His analysis took into account

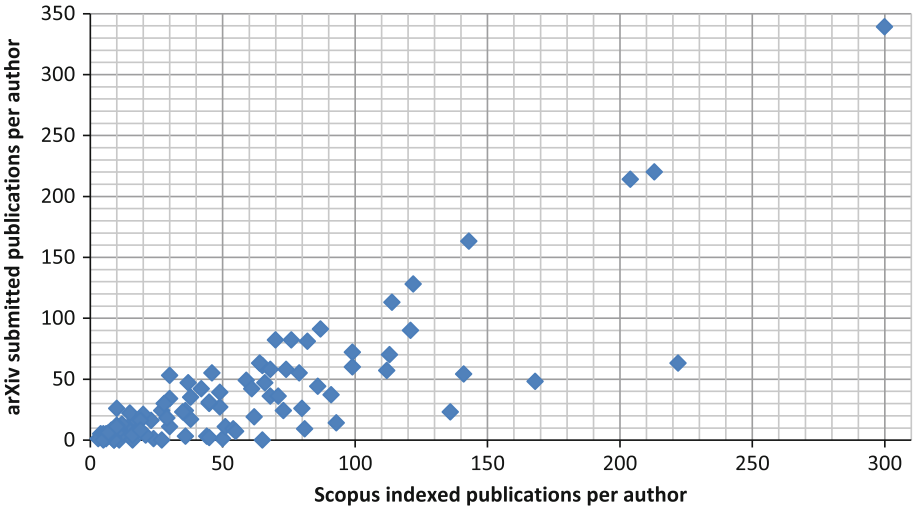

Fig. 1 Number of Scopus indexed items versus number of items submitted to arXiv per author ($R^2 = 0.655$)

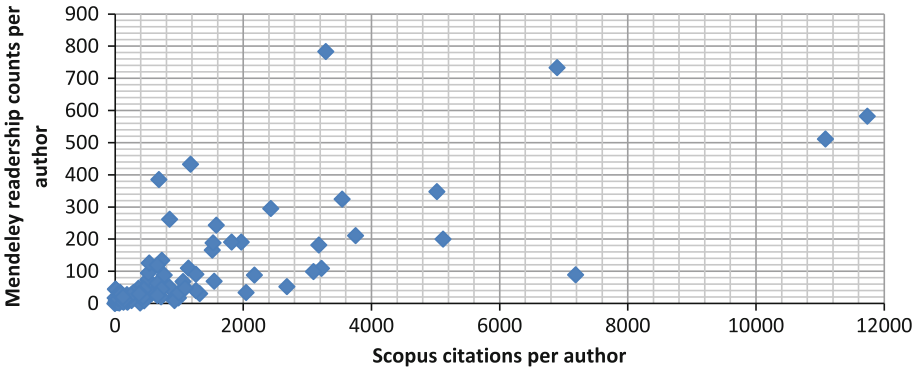

Fig. 2 Scopus citations per author versus Mendeley readership counts per author for Scopus indexed items ($R^2 = 0.517$)

coauthorship. Here we only compared average number of citations received by Scopus indexed items that were submitted to arXiv, versus items not submitted to arXiv. For 90 researchers out of the 100, there were some Scopus indexed items that were not submitted to arXiv (based on title matching), and there were some items that were submitted to arXiv, i.e. we were able to partition their Scopus indexed publications into two non-empty subsets. For each of these 90 researchers we computed the average number of citations received from Scopus by the publications that were submitted to arXiv and indexed by Scopus (16.34 citations) with the citations received by publications that were not submitted to arXiv and indexed by Scopus (17.37 citations), thus we observed a slight difference in the average number of citations in favor of Scopus indexed items, that were not submitted to arXiv, however the difference was not significant. In addition for 57 out of the 90 researchers (63 %) the average number of citations received by items submitted to arXiv was higher than for the items not found in arXiv, partially supporting Moed's findings.

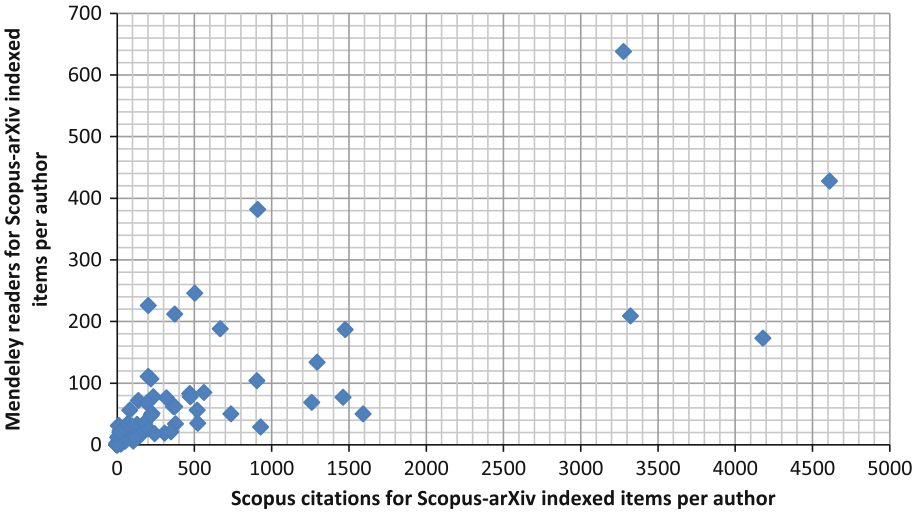

Fig. 3 Scopus citations per author versus Mendeley readership counts per author for Scopus and arXiv indexed items ($R^2 = 0542$)

There were seven researchers in the sample for whom no submissions were located in arXiv. The largest number of submissions was 339 items authored by Gabrielle Ghisellini, a full professor at the Astronomical Observatory of Brera in Italy. His name appears on the list of Thomson–Reuters highly cited researchers (http://highlycited.com/). The number of publications indexed by Scopus at the time of data collection was 300, and the overlap between arXiv and Scopus was 224. The total number of citations received by the 300 indexed items was 11,085. The most cited item was cited 1,032 times; this article was published in 2004 in the Astrophysical Journal, entitled "The *Swift* Gamma-Ray Burst Mission" and was coauthored by 71 researchers. Out of the items appearing either in arXiv or in Scopus (415 items), 203 were bookmarked in Mendeley and their total readership count was 511. The most read item coincided with the most highly cited item and had 23 readers. Gabrielle Ghisellini also had the highest number of indexed items in Scopus among the sample. The lowest number of Scopus indexed items was 2, by a female postdoctoral researcher, and her three publications received only one citation at the time of data collection. Both her publications were dated 2011. The researcher whose publications received the highest number of citations (11,741) was Andrew R, Liddle, a full professor from the University of Edinburgh. He had 204 items indexed by Scopus and 214 submitted to arXiv. He obviously considers arXiv a comprehensive data source, because on his homepage (http://astronomy.sussex.ac.uk/∼andrewl/), instead of providing a list of publications he links to searches for his name on arXiv. The highest total readership count of arXiv submitted publications were authored by Sabino Matarrese (808 readers), and the highest total readership count of Scopus indexed publications were authored by Andreas Freise (783 readers). Andreas Freise is a reader at the University of Birmingham, and had five papers with between 20 and thirty readers each, so the high total number of readers was not the result of a single "hot" paper in this case. Sabino Matarrese is a professor at the University of Padova, and this case too, the high total readership was a result of many papers (135) bookmarked on Mendeley, and not because of a few very highly read papers. Note that it could easily be the case that 30–40 users of Mendeley are interested in the

researcher's work and systematically try to read all his publications. On the other hand it is also possible, although not very probably, that each paper is read by a different group of users.

The most highly cited paper in the sample was a review article on particle physics, cited 3590 times. It was coauthored by Andrew R. Liddle and 172 other authors, published in 2008 in Physics Letters E. This article had 25 Mendeley readers. The article with the highest number of readers (46 readers) was an article published in Nature in 2011, by Saskia Hekker and 25 coauthors on "Kepler detected gravity-mode period spacings in a red giant star". At the time of data collection it was cited 17 times according to Scopus. Saskia Hekker is a postdoctoral researcher at the University of Amsterdam. It should be noted that hyperauthorship (Cronin 2001; Milojevič 2010) is quite common in astrophysics.

Conclusions

In this paper we took a close look at 100 astrophysicists, and compared three data sources: arXiv, Scopus and Mendeley. We found that even though submitting to arXiv is believed to be the norm in astrophysics, Scopus indexes more items than arXiv: there were 82 authors out of the 100 for whom more records were found in Scopus than in arXiv.

The readership counts of the astrophysics articles in the sample are much lower than the citation counts on average, but the two are highly and significantly correlated at the author level.

We cannot generalize from this sample, but the findings are supported by the findings of a much larger study (Larivière et al. 2013). The smaller size of the study allowed us to study some characteristics of the articles and the authors in the sample. Matching between the different data sources was based on authors and titles, and was carried out semi-automatically, thus it is quite possible that the data are not error-free, although precautions were taken to filter out mistakes and to improve the accuracy of the dataset.

An interesting future direction, suggested by one of the reviewers of this paper would be to try to explore why some astrophysicists do not submit their work or some of their work to arXiv.

Acknowledgments This research reported in this paper was funded by the EU FP7 ACUMEN project (Grant Agreement: 266632).

References

arXiv. (2013a). *Wikipedia, The free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=arXiv&oldid=565824986.

arXiv. (2013b). *arXiv submission rate statistics*. Retrieved from http://arXiv.org/help/stats/2012_by_area/index.

Bar-Ilan, J. (2012a). *JASIST@Mendeley*. Presented at the *altmetrics12 workshop of the ACM Web Science Conference*. Retrieved from http://altmetrics.org/altmetrics12/bar-ilan/.

Bar-Ilan, J. (2012b). JASIST 2001–2010. *Bulletin of the American Society for Information Science and Technology, 38*(6), 24–28.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, S., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. In Proceedings of the 17th international conference of science and technology indicators, Montreal, Canada, (pp.98–109).

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*(7), 558–569.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4):e123. Retrieved from http://www.jmir.org/2011/4/e123/.

Grinsparg, P. (1994). First steps towards electronic research communication. *Computers in Physics, 8*(4), 390–396.

Henning, V., & Reichelt, J. (2008). Mendeley—A Last.fm for research? In *Proceedings of the Fourth IEEE International Conference on eScience*, pp. 327–328.

Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology, 44*, 1–64.

Larivière, V., Macaluso, B., Sugimoto, C. R., Milojevic, S., Cronin, B., & Thelwall, M. (2013). The nuanced nature of e-print use: A case study of arXiv. In *Proceedings of the 14th international society of scientometrics and informetrics conference*, Vienna, Austria, vol II, (pp.1321).

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In *Proceedings of the 17th International conference of science and technology indicators*, Montreal, Canada, (pp.451–551).

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics, 91*(2), 461–471.

Mendeley. (2013). *Wikipedia, The free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Mendeley&oldid=568824151.

Milojevič, S. (2010). Modes of collaboration in modern science: beyond power laws and preferential attachment? *Journal of the American Society for Information Science and Technology, 61*(7), 1410–1423.

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of arXiv's condensed matter section. *Journal of the American Society for Information Science and Technology, 58*(13), 2047–2054.

Mohammadi, E., & Thelwall, M. (in press) Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the American Society for Information Science and Technology*.

Priem, J., Piwowar, H., & Hemminger, B. (2012). *Altmetrics in the wild: Using social media to explore scholarly impact*. Retrieved from http://arXiv.org/html/1203.4745v1.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto/.

Schneider, J. W. (2013). Caveats for using statistical significance test is research assessments. *Journal of Informetrics, 7*(1), 50–62.

Shema, H., Bar-Ilan, J., & Thelwall, M. (in press). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source of alternative metrics. *Journal of the American Society for Information Science and Technology.*

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. (2013). Do altmetrics work? Twitter and ten other candidates. *PLoS ONE, 8*(5), e64841.

Wouters, P., Costas, R. (2012). *Users, narcissism and control—Tracking the impact of scholarly publications in the 21st century.* Utrecht: SURF foundation. Retrieved from: http://www.surffoundation.nl/nl/publicaties/Documents/Users%20narcissism%20and%20control.pdf.

Zahedi, Z., Costas, R., & Wouters, P. (2013). How well developed are altmetrics? Cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. In Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria, vol. I, pp. 876–884.