Received: 4 June 2019 | Revised: 27 September 2019 | Accepted: 7 November 2019

DOI: 10.1002/asi.24326

RESEARCH ARTICLE

JASIST WILEY

Scientific Journals Still Matter in the Era of Academic Search Engines and Preprint Archives

Lanu Kim[1] | Jason H. Portenoy[2] | Jevin D. West[2] | Katherine W. Stovel[3]

[1]Graduate School of Education, Stanford University, Stanford, California

[2]Information school, University of Washington, Seattle, Washington

[3]Department of Sociology, University of Washington, Seattle, Washington

Correspondence
Lanu Kim, Graduate School of Education, Stanford University, 520 Galvez Mall, Stanford, CA 94304.
Email: lanu@stanford.edu

Funding information
National Science Foundation, Grant/Award Number: 1735194

Abstract

Journals play a critical role in the scientific process because they evaluate the quality of incoming papers and offer an organizing filter for search. However, the role of journals has been called into question because new preprint archives and academic search engines make it easier to find articles independent of the journals that publish them. Research on this issue is complicated by the deeply confounded relationship between article quality and journal reputation. We present an innovative proxy for individual article quality that is divorced from the journal's reputation or impact factor: the number of citations to preprints posted on arXiv.org. Using this measure to study three subfields of physics that were early adopters of arXiv, we show that prior estimates of the effect of journal reputation on an individual article's impact (measured by citations) are likely inflated. While we find that higher-quality preprints in these subfields are now less likely to be published in journals compared to prior years, we find little systematic evidence that the role of journal reputation on article performance has declined.

1 | INTRODUCTION

The number of scientific articles published in a year has roughly doubled every 9 years since the beginning of modern science (Bornmann & Mutz, 2015). For much of this time, scientists have navigated this ever-expanding space using peer-reviewed scientific journals as an organizing and credentialing system: journals typically have a scientific focus and a journal's reputation sends important signals about the merit of the articles it publishes.

Academic journals' importance for scientific advance rests partially on their gatekeeping role, whereby submitted research is evaluated via peer review and a hierarchical editorial process. Various factors, including the rigor of the review process, produce a status ranking of journals that vary in prominence, prestige, visibility, or impact. While there are many ways to operationalize this ranking (Bergstrom, 2007; Garfield, 2006; West, Bergstrom, & Bergstrom, 2010b), all commonly used measures of journal status favor journals whose articles receive, on average, more citations.

Because a journal's status calibrates an article's quality and significance, publication in a high-status journal, then, is a positive predictor of a single article's subsequent impact, measured typically by the article's own citation count (Bornmann & Leydesdorff, 2017; Didegah & Thelwall, 2013; Onodera & Yoshikane, 2015; Tahamtan, Afshar, & Ahamdzadeh, 2016). The impact of the status of the journals in which a scientist publishes may extend even further, into decisions affecting the authors' hiring, promotion, and access to research funding. This self-reinforcing process represents a classic Matthew effect (Merton, 1968), and it is no surprise that several commentators have issued caveats on the perils of overreliance on journal status in academic decision making (Brembs, Button, & Munafò, 2013; Hicks, Wouters, Waltman, Rijcke, & Rafols, 2015; Lariviere et al., 2016; Martin, 2016; Seglen, 1997; Stephan, Veugelers, & Wang, 2017; West, Bergstrom, & Bergstrom, 2010a).

Scholarship published outside of high-status journals has traditionally been more difficult to discover. Digitization has made the process of accessing known articles far easier, but it initially did little to improve search, and the curatorial role of journals persisted. In the past decade, however, online academic search engines as well as online preprint repositories have emerged that are changing the way scientists follow and search for scientific research. Searches conducted with algorithmically driven tools like Google Scholar return lists of articles organized by topic, author, keyword, and prominence, with results weighted in unknowable ways, while online repositories allow scholars to bypass journals altogether. With these new technologies, therefore, an article's visibility in various electronic archives is now at least partially decoupled from the journal's reputation, and articles published in lower-tier journals may have new opportunities to reach an audience.

In light of these recent changes in both search and access, it is an open question whether academic journals will retain their traditional gatekeeping role going forward. Some have argued “no,” pointing to a series of recent reports claiming to find a decline in the effect of elite journals on individual articles' subsequent citation (Acharya et al., 2014; Larivière, Lozano, & Gingras, 2014; Lozano, Larivière, & Gingras, 2012).

However, correctly estimating the independent effect of a journal's influence on article performance is methodologically challenging, since academic articles are nested in journals: “better” articles are more likely to survive the peer-review process at higher-status journals, and yet once an article is published in a journal, readers cannot help but evaluate the quality of the article in light of the journal's reputation. It is therefore almost impossible to separate the effect of journal influence from an individual article's quality in the naturally occurring observational data used in most bibliometric research, and as a consequence, all previous efforts to systematically estimate the magnitude of the effect of journal status on individual article performance have failed to effectively purge the effect of article quality from the “journal effect.”

Here we report significant progress in addressing this methodological problem by using an innovative measure of article quality that is arguably independent of journal influence: citations to articles posted to the arXiv preprint repository before publication. Working articles and preprints have long been used to disseminate research findings before peer review is complete (Brown, 1999; Kreitz, Addis, Galic, & Johnson, 1997). In 1991, arXiv (arXiv.org) was established as an online central repository and clearinghouse for client-side rendering of scientific articles in physics. ArXiv submissions were moderated for topic but not peer-reviewed. Physicists quickly adapted to arXiv, routinely uploading research articles as soon as the article was complete (Brown, 2001; Larivière, Sugimoto et al., 2014). Neighboring disciplines that also used the compact TeX file format, including mathematics, astronomy, computer science, quantitative biology, and statistics, followed physics' lead and soon began relying on arXiv; more recently, several other fields have established similar services (for example, bioRxiv.org launched in 2013; SocArXiv in 2016; ChemRxiv.org in 2017). With the emergence of arXiv as an easy to access source of cutting edge science, researchers' reliance on preprints as a channel of communication increased (Brown, 2001), and citations to unpublished preprints become more common (Noruzi, 2016). Nevertheless, many articles posted to arXiv in physics and other fields are subsequently published in peer-reviewed journals (Larivière et al., 2014).

In this study we exploited the fact that in three subfields of physics—high energy physics phenomenology (Hep-ph), astrophysics, and condensed matter—use of arXiv has been part of scientific practice for several decades. For these early adopters of arXiv, almost all articles are posted to arXiv. Scholars in these subfields make use of arXiv's automated alert systems to stay abreast of new developments, and they regularly cite arXiv preprints. We used these subfields with relatively long histories of using the arXiv as a case-study of how journals' role in science may have changed in response to new technologies for academic search and publications. Our approach is an advance over prior studies because we can treat the number of citations an article in arXiv receives as a proxy for article quality that is independent of journal status (independent because these citations are made before the article has published and therefore are not correlated with the status of any journal). Having this independent proxy for article quality allowed us to generate better estimates of the direct impact of journal influence on an individual article's performance, and assess whether this impact has changed over time.

Using the number of citations an article in arXiv receives before publication as an independent indicator of measure of article quality, we conducted two sets of analyses that addressed three interrelated questions. First, we set out to determine the magnitude of the effect of journal influence on citations after controlling for preprint citations. Even though the distribution of citations to articles is highly skewed (even for articles published in the same journal, Lariviere et al., 2016), measures of journal status such as the Journal Impact Factor (JIF) have been shown to be powerful predictors of citations (Bornmann & Leydesdorff, 2017; Didegah & Thelwall, 2013; Onodera & Yoshikane, 2015). We investigated whether the effect of journal influence holds even after controlling for preprint citations.

Using the same modeling framework and our method of isolating the journal-specific effect, we then investigated whether there are temporal changes in the effect of journal status on article visibility and citation in the era of modern electronic technologies. This analysis provided a test of the hypothesis that the development of new electronic technologies has reduced the influence of the journal (Acharya et al., 2014; Larivière, Lozano et al., 2014; Lozano et al., 2012).

Our second analysis addressed whether there has been a change in the relative quality of articles published in journals. In fields in which new and important research findings are routinely posted to arXiv (and therefore easily located by appropriate audiences), some scientists may skip the journal review process altogether. A previous report (Gentil-Beccot, Mele, & Brooks, 2010) suggested the existence of a "quality bias" in arXiv (better articles and high-impact authors are more likely to be uploaded in the first place); we therefore systematically tested for adverse selection into journals among articles posted to arXiv. Taken together, these analyses provide new insight into the role of scientific journals during a time of rapid technological change.

The results presented here show that, whereas prior estimates of the effect of the journal's reputation on an individual article's citation performance are likely inflated, there is no systematic evidence that the role of journal reputation has declined with the advent of academic search engines. After adjusting for prepublication citation levels, however, we find that higher-performing articles posted to arXiv have, over time, become less likely to be published in journals compared to prior years.

2 | DATA AND METHODS

2.1 | Data

Our analyses were conducted on data from three subfields of physics: high energy physics phenomenology (Hep-ph), astrophysics, and condensed matter. Scientists in each of these fields were early adopters and continuous users of arXiv (ArXiv, 2017; Larivière et al., 2014), and these fields have been studied in previous research on arXiv (Gentil-Beccot et al., 2010; Mine, 2009; Moed, 2007). From arXiv.org we collected basic bibliographic and identifying data about all articles uploaded to arXiv between 1998 and 2013 in each of the three fields.

Using the article's DOI and title, we linked each article to its corresponding entry in the Microsoft Academic Graph (Sinha et al., 2015). We dropped articles that had incomplete information regarding title, author, and/or published year. Details about the data linking process is described in Section A in Appendix S1. From the Microsoft Academic Graph data, we recorded whether, when, and where the article was subsequently published, and we collected citations that went to the arXiv and to published versions of the article. We treat citations to the arXiv preprint as a journal-independent measure of article performance, and citations to the published version a measure of article performance that may be influenced by the journal's reputation. Our final data set included a total of 31,623 articles in Hep-ph, 42,063 in astrophysics, and 84,326 in condensed matter.[1]

The temporal trend in the yearly total count of articles posted to arXiv and citations to arXiv preprints for our three subfields is shown in Figure 1. We found a rapid increase in the total number of submitted articles in astrophysics and condensed matter between 1998 and 2013, but a relatively slower increase in Hep-ph. Citation rates increased following publication rates, but the speed of increase also varied by subfield. Because some researchers upload articles to arXiv simultaneously with journal submission, the yearly count of articles uploaded is larger than the yearly count of citations made to articles in arXiv. When articles are published quickly after posting, there is little opportunity for the arXiv version to

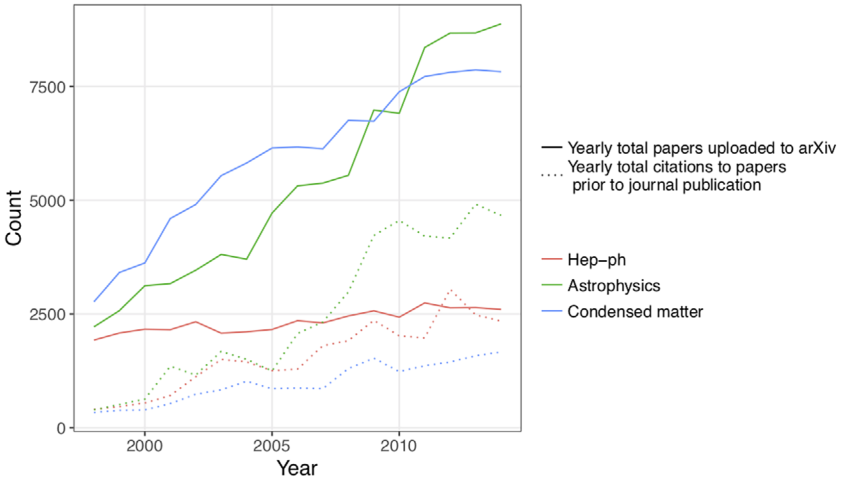

FIGURE 1 The rise of arXiv. Yearly total articles uploaded to arXiv (solid line) and yearly total citations made to arXiv articles prior to journal publication (dotted line) between 1998 and 2014 [Color figure can be viewed at wileyonlinelibrary.com]

be cited. In condensed matter, 92% of articles published within 6 months of posting receive no citations before journal publication. This percentage is 85% for astrophysics and 76% for Hep-ph.

2.2 | Analysis 1: Improving estimates of the impact of journal influence by adding a measure of article quality and reexamining the temporal change

Our first analyses aimed to answer the first two of our three research questions by improving estimates of the direct impact of journal influence on article performance while controlling for article quality and retesting the influence of journal status on article's subsequent citation count. We restricted our data set to only those articles that were posted to arXiv and subsequently published in journals between 1998 and 2013. To measure the article's performance after journal publication, we counted the number of citations the article received in the 3 years after publication.

For the statistical analysis, we used a zero-inflated regression model with a negative binomial distribution (Didegah & Thelwall, 2013; Chen, 2012). A large fraction of articles received no citations, and so we model the overall citation distribution with the zero-inflated regression model because it treats the process of generating zero values as a mixture of two processes: first, an author might not cite an article because it is not relevant to their research; and second, authors' decisions about what articles to cite (or not cite) reflect multiple and heterogeneous decision rules and conventions (Garfield, 1965; Tahamtan & Bornmann, 2018; Wang & Domas White, 1999).[2] Because the citation distribution shows a positively skewed shape, we applied a negative binomial distribution.[3] To account for within-journal correlation, we used robust standard errors clustered by journals. Further information can be found in the detailed model results in Section C in Appendix S1.

Our key explanatory variable in this model is the journal's ArticleInfluence (AI) score (West et al., 2010b), which is a network-based, journal-level measure of influence that is normalized by the size of the citing journal. AI is based on the Eigenfactor (West et al., 2010b), a variant of PageRank, in which citations from high-status journals receive more weight than citations from more peripheral journals. While we used AI as a measure of journal influence, we use the term "journal influence" in the main text and Appendix S1 to avoid possible confusion of this measure as an article-level metric.

In addition to our measure of journal influence, our models included the article's prepublication citation count—our journal-independent measure of article performance. In treating the prepublication citation count as an indicator of article quality, we are assuming that the published articles were not substantially revised since the initial submission to arXiv, an assumption that is consistent with a recent empirical study that showed "the text contents of ... scientific articles generally changed very little from their pre-print to final published versions" (Klein, Broadwell, Farb, & Grappone, 2018, p. 1).[4] To capture the time trend, we included a variable measuring the number of years since our base year, 1998.[5] Finally, we also controlled for the number of months since the date an article was first posted in arXiv.

With this modeling framework, we tested three nested models. The first model (Model 1 in Section C in Appendix S1) included only the journal influence and two basic control variables: the number of months the article appeared in arXiv before journal publication and the number of years since 1998. The second model (Model 2 in Section C in Appendix S1) added our key control variable: the article's prepublication citation count while in arXiv. In the third model (Model 3 in Section C in Appendix S1), we added an interaction effect of journal influence and the number of years since 1998, thereby allowing us to assess our second research question: Has the impact of journal influence changed over time?

2.3 | Analysis 2: Temporal change in the quality of articles that are published in a journal

For the third research question, we asked whether there has been a change over time in the relative quality of articles from arXiv that are subsequently published in journals. Our aim here was to determine if highly cited arXiv articles are now less or more likely to appear in a journal. We used the Cox's proportional hazards regression model (survival analysis, Cox & Oakes, 1984) to analyze time-to-publication for articles published in arXiv in each of the three fields, adjusting for fixed and time-varying covariates. Time to publication was measured by following articles uploaded to arXiv between 1998 and 2014 for 24 months or until the article was published, whichever came first.[6] Because the data showed a relatively constant increase in articles uploaded to arXiv and published in a journal through 2016, the last publication year we used in the analysis was 2014, and the last citation was made in December 2016. The primary explanatory variable in these models is a time-varying measure, the square root of the current cumulative citation count to the arXiv article by month t, which represents the currently observed quality of the article. We chose the square root rather than natural log transformation due to the large number of zeros values. In addition to the cumulative citation count, we included the number of years since 1998, where the year is when an article is first uploaded to arXiv. Finally, we interacted the article-specific cumulative citation measure with the number of years since 1998, in order to evaluate whether the relationship between article quality and time-to-publication has changed over time.

3 | RESULTS

3.1 | Effect of journal influence on article performance

Figure 2 summarizes the effect of journal status on a published article's predicted citation count, with and without adjustment for prior citation to the article in arXiv (our journal-independent measure of article performance) for each of the three subfields studied. Consistent with prior studies (Bornmann & Leydesdorff, 2017; Didegah & Thelwall, 2013; Larivière & Gingras, 2010; Onodera & Yoshikane, 2015), we found that in all fields citations to the published article increased with journal influence; however, these effects were attenuated when the journal-independent measure was considered. Since the model fit significantly improved after controlling for the article's prior arXiv citation count (see Section C in Appendix S1), this measure of article quality improves the estimates of articles' post-publication performance. However, the coefficient of journal effect decreases after adding the new measure. The improved model shows that differences in preprint citations make a substantive difference in predicted citation counts for articles ultimately published in the same journal. For articles that received no preprint citations, the naïve model overestimated subsequent citations: for example, in journals with a journal influence score of 2, articles with zero citations were predicted to have 12–38% more citations than the prediction from the model controlling for article quality. In contrast, the more an article was cited before journal publication, the more the naïve model underestimated predicted citation counts: If an article received two citations before journal publication, the model that failed to control for article quality predicts 18–54% fewer citations. The degree to which citations were underestimated is even greater for articles whose arXiv version was cited more often: Postpublication citations to articles with five prepublication citations were underestimated by 62–88%.

3.2 | Temporal changes in journal impact on subsequently received citations

We next investigated whether the general growth in the use of the arXiv is associated with a change over time in the effect of journal impact on citations to published

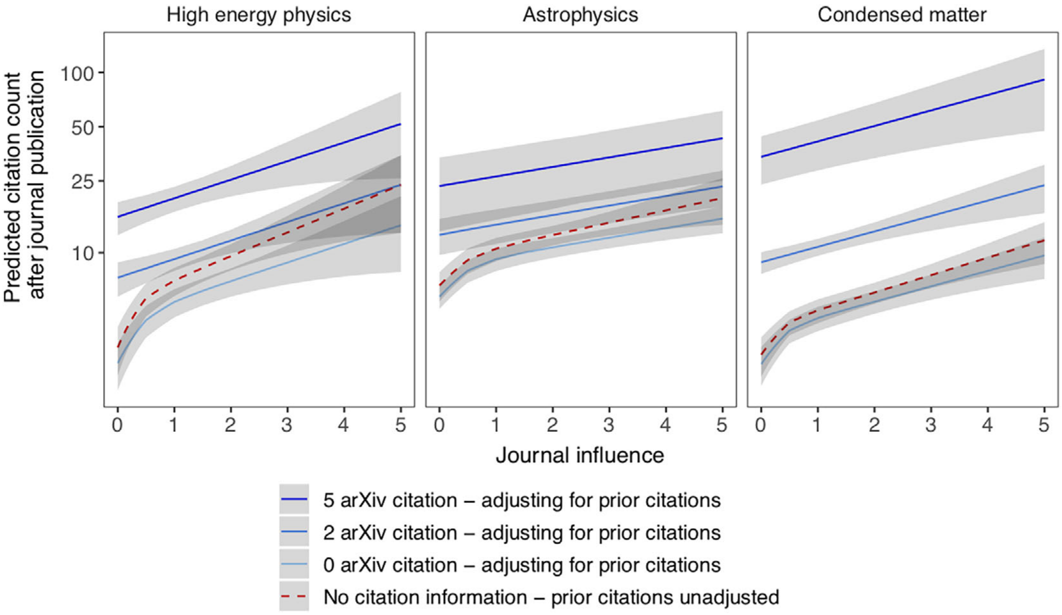

FIGURE 2 Preprint citation count information improves prediction of article performance over models using only journal influence. The effect of journal influence on the predicted citation count accumulated in the 3 years after journal publication. Red dotted line is from the model without controlling arXiv preprint citations, and three blue lines are from the model after controlling preprint citations (from the brightest, receiving 0 prior arXiv citations, 1 citation, and 5 citations). The shaded areas represent the 95% confidence interval. Model details are included in Section C (Model 1 and 2) in Appendix S1 [Color figure can be viewed at wileyonlinelibrary.com]

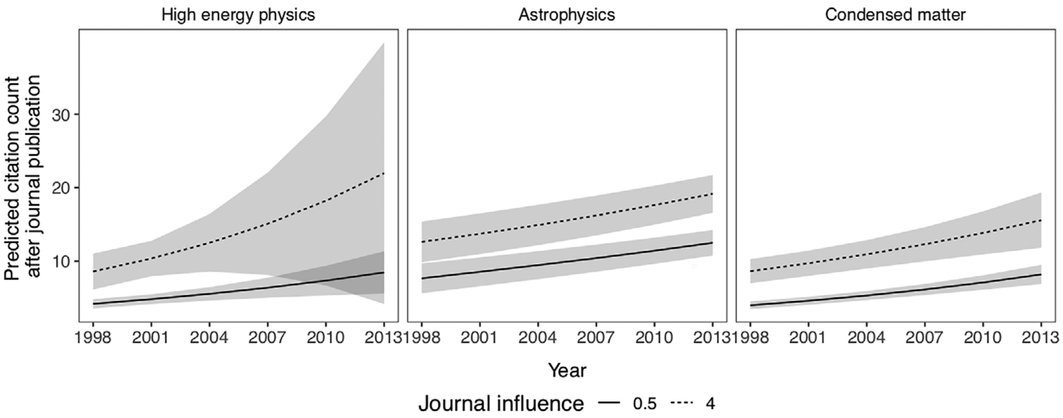

FIGURE 3 No decrease in journal influence found. The effect of time on the predicted citation count accumulated in the 3 years after journal publication by two journal influence scores (0.5 and 4; line types indicate journal influence). Lines are calculated from Model 3, which estimates the interaction effect of time and journal influence on citation count, adjusting for arXiv citations. Shaded areas represent the 95% confidence interval. Model details in Section C (Model 3) in Appendix S1

articles. Figure 3 reports, for each subfield, the average predicted citation count for articles published in journals with higher and lower influence scores, adjusting for preprint citations. Parallel lines indicate that the citation gap between two journals neither decreases nor increases over time, which suggests no temporal change in the effect of journal influence on article performance. In Hep-ph, the citation gap between higher- and lower-status journals has actually grown over time, with articles published in higher-status journals now receiving even more of a citation benefit than in earlier years. However, the confidence intervals of the predicted values in later years are too broad to conclude that there has actually been an increase in the effect of journal influence. In astrophysics and condensed matter, the citation gap between higher- and lower-status journals has not changed over time. Combining these results, we found that after controlling for preprint citations there is no evidence of a decrease in the effect of journal status on articles' postpublication performance.

3.3 | Effect of article quality on journal publication

Figure 4 shows, for each subfield, the effect of article quality (measured as the article's cumulative citation count) on journal publication, and how this may have changed over time. In all three cases, the marginal effect of article quality on publication rates decreased over time but still remained positive. This means that as time goes forward, arXiv articles with many citations were less likely to be published in a journal. These models provide evidence that in each of these subfields, better-performing articles (measured via citations to the preprint) are now less likely to be published in a journal compared to prior years.

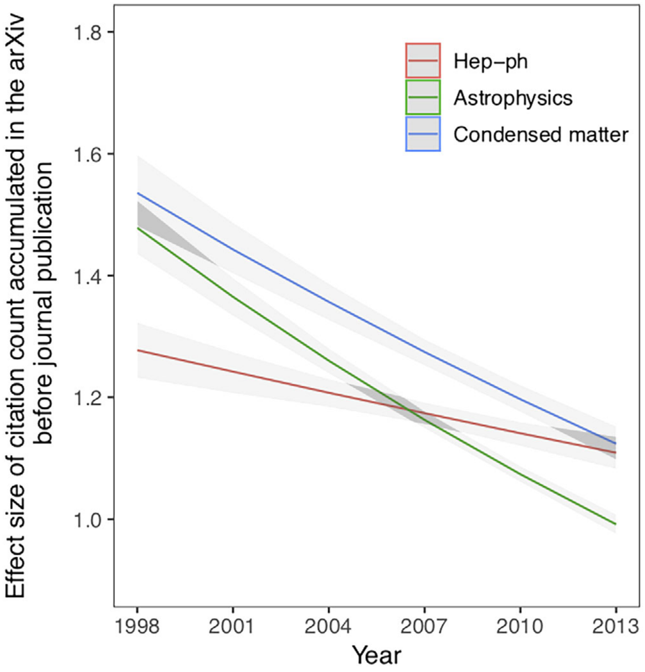

FIGURE 4 Over time, articles with more preprint citations are less likely to be published in journals. The effect of cumulative citation to all arXiv articles on the hazard of journal publication, by subfield. Shaded areas represent the 95% confidence interval. Model details in Table S5 in Appendix S1 [Color figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

In this study we investigated possible changes in the extent to which peer-reviewed journals serve as a credentialing system for scientific research. We began by introducing a novel measure of article quality—citations to the article in arXiv—that allowed us to estimate the influence of journal status on a published article's performance, measured via citations, more accurately. Including this new measure in models of citation counts allowed us to isolate the independent impact of journal status more effectively, and to assess whether this effect has changed over time. We also investigated whether the temporal trend is associated with changes in the quality of articles submitted for journal publication.

Until very recently, scientific articles were most often read and cited by readers who knew where the article was published. Conference presentations and hard-copy preprints provided opportunities for close colleagues to know of each others' work, but these distribution channels are limited in scope. Thus, in evaluating the significance of a published article, scientists were forced to rely in part on the reputation of the journal that published it, a reliance that confounds intrinsic article value with characteristics of the journal and affects the article's subsequent visibility and impact. Drawing on newly available data we are able to provide a more precise quantification of the journal effect by controlling for a measure of article quality that we constructed by linking data from arXiv with publication and citation data from the Microsoft Academic Graph. Readers of preprint articles on arXiv do not typically know in which journals they will be published in the future; thus, accumulated citation data to preprint versions of an article can be used as an indicator of article quality that is independent of journal influence. After introducing this control for article quality, we found that the impact of journal influence is lower than in models that do not control for article quality. Nevertheless, publication in higher-status journals is still associated with increased citations.

Our study of temporal changes in the effect of journal influence on an article's citation count revealed that after controlling for article quality, there is no evidence of a decrease in the importance of journal status on an article's citation count, even in fields with heavy reliance on arXiv. In the three subfields we analyzed, the impact of journal influence had not declined at all. Thus, concerns about the “demise of journals” may be overblown, at least in these fields.

In answering our third question, we found that while articles receiving more attention in arXiv are more likely to be published, in all three subfields the effect decreased over time. This finding suggests that as arXiv and preprints in general become more popular, authors may be less inclined to pursue journal publication if their preprint articles are sufficiently acknowledged.

Taken together, we conclude that in fields where research is indexed, read, and cited in new channels like arXiv, journals continue to act as status markers signaling article quality and significance, even as some authors bypass journal publication altogether. Our findings therefore contradict prior work showing a declining impact of elite journals (Acharya et al., 2014; Larivière, Lozano et al., 2014; Lozano et al., 2012).

The generalizability of our results is limited at present. The particular subfields of physics we analyzed are less interdisciplinary than many other scientific disciplines (Van Noorden, 2015), and therefore scientists in these fields might behave differently than in fields that frequently bridge areas of scientific expertise. Thus, our findings might not characterize less highly specialized and fast-moving fields. Because arXiv initially targeted a small number of highly technical scientific fields, comparable and solid longitudinal data for more interdisciplinary fields such as social sciences are not available at present, so we could not evaluate whether there are differences between fields. However, we believe our analyses of these early-adopting fields may help us glimpse the future of other disciplines that are currently undergoing similar changes in their own publication environments. The recent burst of preprint services in the life sciences and social sciences will likely offer the opportunity to conduct a similar analysis in these fields.

A second potential limitation of our study is that we measured the quality of an article by counting how many citations it received. Although highly cited articles often correspond with what researchers perceive as the impactful research (Garfield, 1999), scholars cite an article for various reasons, and therefore more citations do not necessarily indicate the higher “quality” of an article. For example, researchers often pay close attention to articles written by high-status authors (Simcoe & Waguespack, 2011), and the subsequent citations serve to amplify an authors' existing status, bringing even more citations to already successful articles. However, because it naturally controls for all author- and article-level characteristics, our use of this measure has merit because it isolates the unique effect of the journal impact factor on an article's subsequent citations (Garfield, 1999; Hicks et al., 2015).

As with most empirical studies, there are possible sources of selection bias in our data. The life cycle of a research article includes many stages, and our analysis plan was designed to capture an article's progress through these increasingly selective stages. Among articles uploaded to arXiv, our approach traces which of them are published in a journal, and how many citations those articles receive after journal publication. Nevertheless, our research design is exposed to two potential sources of selection bias. First, our data source itself may be selective, because we were not able to link all arXiv articles to entries in Microsoft Academic Graph.

However, as far as we are aware, Microsoft Academic Graph attempts to index all articles from arXiv, and thus we feel reasonably confident that we can treat unlinked articles as missing at random. A potentially greater source of selection bias comes from the fact that we did not analyze articles that are published in a journal but never uploaded in arXiv. We excluded these articles because we do not systematically observe our measure of article quality from them. Nonetheless, we do not believe that this exclusion undermines our key finding that the impact of journals has not declined appreciably over time. Published articles that were deposited in arXiv receive more citations because free preprint services disseminate research to potential readers (Gentil-Beccot et al., 2010). If the only channel that readers could use to access articles was through journals, the impact of a journal should be higher on average, not lower.

5 | CONCLUSION

Our results suggest that journals remain an important status marker for scientific work despite the advent of preprint archives and algorithmic search technologies. As a consequence, journals in the fields we examined continue to play an outsized role not only in communicating scientific results, but also in validating scientists' scholarly achievements. Such validation is especially important for those early in their career who seek to pass through critical milestones including jobs, tenure, awards, and grants. And yet even as the signaling effect of journals is needed more than ever, the peer review system is being strained with more and more articles being submitted, leading to lower-quality reviews. The review process at many journals has become more competitive and in some cases drawn-out (Powell, 2016). As a result, established researchers who no longer need validation for their career may choose to skip the bothersome review process and go exclusively to preprint services like arXiv that facilitate the dissemination of research findings with no claims about the quality of uploaded works. This interpretation is supported by our empirical finding that highly cited arXiv articles are now less likely to appear in journals than was the case when arXiv was new. Future research that stratifies publication patterns by scientific generation will allow us to further untangle journals' communication and validation functions in this era of technological change.

ACKNOWLEDGMENTS

This research was supported by NSF award 1735194, awarded to Katherine Stovel and Jevin West. The authors thank Sam Wang and Alex Wade for helpful conversations.

ORCID

Lanu Kim https://orcid.org/0000-0002-7381-4959

ENDNOTES

[1] The code and data used in this article can be found at the following link: https://github.com/lanukim/arXiv-article-data-and-code.

[2] An alternative to the zero-inflated model is a hurdle model, which specifies one process for zero counts and another process for positive counts. A zero-inflated regression model has an advantage over a hurdle model because it assumes that noncited articles can be generated from either of two processes, while a hurdle model assumes that zeros are only generated from one process. Thus, we believe a zero-inflated model better matches the processes that produce uncited articles in science.

[3] We make this decision to allow overdispersion. The coefficients of the overdispersion parameter (the natural log of alpha in Stata software) for each of the three subfields we analyze were statistically significant with a .05 alpha level in many models, which supports the necessity of using a negative binomial distribution.

[4] Because articles published shortly after being uploaded to arXiv might have had not enough time to be widely read, we checked the robustness of our results by removing articles that were in arXiv for 6 months or less before they appeared in a journal. The results are similar to those derived from our original data set (Section B in Appendix S1).

[5] However, Figure 1 illustrates that while the use of arXiv in the three subfields has risen in all cases, it has done so at different rates. Thus, we did a robust check by measuring temporal trends with the total attention received by articles in arXiv in a subfield and in year *y*. Results from this specification do not substantively differ from our main results.

[6] We stopped tracking (that is, right-censored) articles not published within 24 months of posting to arXiv. A total of 7.1% of the articles in Hep-ph, 3.1% of astrophysics, and 1.8% of condensed matter were not published within 24 months of being posted to arXiv, and are therefore right-censored.

REFERENCES

Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., & Shetty, N. (2014). *Rise of the rest: The growing impact of non-elite journals*. arXiv preprint arXiv:1410.2217.

ArXiv. (2017). *arXiv submission rate statistics*. Retrieved from https://arxiv.org/help/stats/2017_by_area

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, *68*(5), 314–316.

Bornmann, L., & Leydesdorff, L. (2017). Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data. *Journal of Informetrics*, *11*(1), 164–175.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, *7*, 291.

Brown, C. (1999). Information seeking behavior of scientists in the electronic information age: Astronomers, chemists, mathematicians, and physicists. *Journal of the American Society for Information Science*, *50*(10), 929–943.

Brown, C. (2001). The E-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science and Technology*, *52*(3), 187–200.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, *63*(3), 431–449.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. Boca Raton, FL: Chapman and Hall/CRC.

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064.

Garfield, E. (1965). Can citation indexing be automated?. In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation: Symposium proceedings* (pp. 189–192). Washington, DC: National Bureau of Standards.

Garfield, E. (1999). Journal impact factor: A brief review. *Canadian Medical Association Journal*, *161*(8), 979–980.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93.

Gentil-Beccot, A., Mele, S., & Brooks, T. (2010). Citing and reading behaviours in high-energy physics. *Scientometrics*, *84*(2), 345–355.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature*, *520*(7548), 429–431.

Klein, M., Broadwell, P., Farb, S.E., & Grappone, T. (2018). *Comparing published scientific journal articles to their pre-print versions—extended version*. arXiv preprint arXiv:1803.09701.

Kreitz, P. A., Addis, L., Galic, H., & Johnson, T. (1997). The virtual library in action: Collaborative international control of high-energy physics pre-prints. *Publishing Research Quarterly*, *13*(2), 24–32.

Lariviére, V., & Gingras, Y. (2010). The impact factor's Matthew Effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology*, *61*(2), 424–427.

Lariviere, V., Kiermer, V., MacCallum, C.J., McNutt, M., Patterson, M., Pulverer, B., … Curry, S. (2016). *A simple proposal for the publication of journal citation distributions*. BioRxiv:062109.

Larivière, V., Lozano, G. A., & Gingras, Y. (2014). Are elite journals declining? *Journal of the Association for Information Science and Technology*, *65*(4), 649–655.

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, *65*(6), 1157–1169.

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*, *63*(11), 2140–2145.

Martin, B. R. (2016). Editors' JIF-boosting stratagems–Which are appropriate and which not? *Research Policy*, *45*(1), 1–7.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.

Mine, S. (2009). The roles and place of arXiv in scholarly communication. *Library and Information Science*, *61*, 25–58.

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.

Noruzi, A. (2016). arXiv popularity from a citation analysis point of view. *Webology*, *13*(2), 1–7.

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

Powell, K. (2016). Does it take too long to publish research? *Nature*, *530*(7589), 148–151.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497–502.

Simcoe, T. S., & Waguespack, D. M. (2011). Status, quality, and attention: What's in a (missing) name? *Management Science*, *57*(2), 274–290.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., & Wang, K. (2015, May). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246). New York: ACM.

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature*, *544*(7651), 411–412.

Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225.

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, *12*(1), 203–216.

Van Noorden, R. (2015). Interdisciplinary research by the numbers. *Nature*, *525*(7569), 306–307.

Wang, P., & Domas White, M. (1999). A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages. *Journal of the American Society for Information Science*, *50*(2), 98–114.

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010a). Response to Big Macs and Eigenfactor scores: The correlation conundrum. *Journal of the American Society for Information Science and Technology*, *61*(12), 2592–2592.

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010b). The Eigenfactor MetricsTM: A network approach to assessing scholarly journals. *College & Research Libraries*, *71*(3), 236–244.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kim L, Portenoy JH, West JD, Stovel KW. Scientific Journals Still Matter in the Era of Academic Search Engines and Preprint Archives. *Journal of the Association for Information Science and Technology*. 2019;1–9. https://doi.org/10.1002/asi.24326