

# Are peer review duration and publication delay research quality signals?

Paul Donner\*

\*[donner@dzhw.eu](mailto:donner@dzhw.eu)

ORCID: 0000-0001-5737-8483

Department 2 Research System and Science Dynamics, German Centre for Higher Education Research and Science Studies (DZHW), Germany

Here we study how the lengths of the periods from submission to acceptance (review duration) and from acceptance to publication (publication delay) relate to research quality, as operationalized by F1000Prime recommendations, for a large dataset of publications from the life and health sciences. We find a statistically detectable relationship between shorter peer review duration and recommendations, but its effect size is negligibly small.

## 1. Introduction

It has long been one of the central methodological concerns of scientometrics to get as close as possible to measuring the construct of research quality with data from the scientific communication system. Using citation counts and derived measures for quantifying scientific impact is the major operationalization of this approach, which is commonly justified on the grounds that citation counts have been found to co-vary to some degree with quality judgments (Aksnes, Piro, & Fossum, 2023; Allen *et al.*, 2009). However, many other factors besides research quality, including social and strategic motivations, also induce authors to cite specific references (Beck *et al.* 2018; Brooks, 1986).

Citation counts may not be the only bibliometrically accessible data that carry some signal of research quality. High quality papers might be recognized in the submission and review process and be treated differently than lower quality papers. If reviewers perceive high manuscript quality they might be motivated to finish their reviews sooner to speed up the dissemination of important findings and because they enjoy engaging with intellectually stimulating research. In addition, papers of high quality might from the outset have less flaws that need to be revised, which would also shorten review duration. Similarly, editors might prioritize papers they assess as being of higher quality and importance. On the other hand, it is possible that unusually important or unconventional papers are scrutinized more carefully and critically, perhaps because of their more controversial claims or greater societal implications, which would lead to longer review durations.

Some indirect support for the first hypothesis comes from studies of the relationship of peer review duration and citation impact. Kousha & Thelwall (2022), for instance, found some evidence that, within journals, Covid-19 papers with shorter review durations were associated with higher citation impact. This was not the case for comparable papers on other topics. Shen *et al.* (2015) studied ‘editorial delay’ (time from submission to acceptance) and citation counts for papers in three highly visible journals. They found some significant correlations for some journal-year combinations but no overall convincing statistical signal. Rigby, Cox, & Julian (2018) analyzed associations between various peer review variables and citation counts for one

business and management journal. They also found an association between short review duration and citation counts in a regression analysis. Zhou *et al.* (2023) found a robust, albeit small ( $r \approx 0.1$ ) association of peer review time to negative sentiment expressed in peer review reports in one large journal.

There is also some evidence supporting the second hypothesis. Hilmer & Lusk (2009) studied various factors influencing citation counts of articles in two agricultural economics journals, including review duration in one of the journals. Regression analysis revealed a significant positive association between longer review duration and eventual citation count.

These findings only indirectly relate to research quality of papers because all studies used citations as the dependent variable which is at best a moderately reliable proxy for quality. The results neither show a robust effect nor clearly rule out either of the two possible directions of the hypothesized relationship and leave open the possibility of a nonlinear relationship, such as an U-shaped functional form where both low-quality and high-quality papers have longer times in review and from acceptance to publication than middle-quality papers.

Here we investigate the relationship between publication quality ratings, more precisely public expert paper recommendations, and peer review duration (the time from submission to acceptance) and publication delay (the time from acceptance to publication), respectively. In addition, we test whether the inclusion of these two variables improves the prediction of paper quality from variables available at publication time. Many studies have shown that individual papers' citation counts some years after publication can be predicted with some accuracy at publication time using such variables as JIFs, numbers of authors, countries, references, pages and characteristics of cited references (e.g. Bornmann & Leydesdorff, 2015; Haslam *et al.*, 2006; Onodera & Yoshikane, 2015). Some of these factors have been shown to correlate with citation counts after controlling for quality ratings, e.g. reference citation performance, but not author count, in Bornmann *et al.* (2012) and JIF and author count in Bornmann & Leydesdorff (2015). However, prediction of explicit recommendation, as a more direct and explicit research quality indicator, from variables available at publication time, has not been studied.

## **2. Data and methods**

For publication quality scores we use ratings from the life and health sciences paper recommendation platform F1000Prime. On F1000Prime (now called Faculty Opinions), which is a subscription-based service, recognized and established researchers can recommend publications they deem important. They can assign a quality rating score of one of three levels ('good', 'very good', 'exceptional') and provide a brief review or comment on the paper to explain its significance. F1000Prime provided our research group with a data extraction in 2017 for research purposes.

Paper submission and acceptance dates are not indexed by major multidisciplinary bibliographic-bibliometric databases but PubMed does index them and also the date of the exact day of first publication, so we use PubMed records for the duration variables. Because of the disciplinary scopes of F1000Prime and PubMed, the study is restricted to the life and health

sciences. PubMed records for ‘article’ publications of publication years 2007–17 were downloaded with Entrez Unix tools and relevant dates and PubMed IDs extracted. Dates were checked for validity and consistency and some irregular records excluded.

The variable ‘review duration’ was calculated as the number of days from submission to acceptance; the variable ‘publication delay’ as the number of days from acceptance to publication. As we only have duration data for published papers and only for the journal they were eventually published in, it is an important limitation of this study that we do not have accurate data for papers which have been rejected from any number of other journals and have been through several rounds of review elsewhere. Therefore, the reliability of the timing variables for reflecting true time under review and revision is unknown, and likely quite low.

PubMed data were matched to Web of Science (WoS) records and F1000Prime ratings by PubMed ID. We included all WoS and PubMed records from journals for which at least one F1000Prime rating was available in the data. This restricts the disciplinary scope of the data set to those disciplines and journals with which F1000Prime members are familiar to some degree. Such a restriction means we still include records for publications without any F1000Prime ratings. This enables us to interpret an absence of any F1000Prime rating as an expression of an additional low quality rating class, ‘unrated’. The advantage of this procedure is to circumvent the strong restriction of range on quality inherent in the F1000Prime system, i.e. its preselection for quality (Waltman & Costas, 2014) as also recommended by Bornmann & Leydesdorff (2015, p. 428). One disadvantage is that this possibly excludes entire journals which are in scope but of too low quality to have attracted F1000Prime recommendations to their papers, so there is still some range restriction. Note also that journals which do not report both submission and acceptance dates are not included.

From the WoS data, a number of variables were calculated which have previously been found to predict later citation counts on the level of individual publications. These will here be used to test if and how much they predict research quality. These following variables were extracted:

- average JIF (average JIF value of the journal for the years 2009–17)
- number of authors and number of different countries involved in the paper
- number of cited references
- subfield of the journal according to ScienceMetrix Journal Classification (SMJC) version 2.0 (Archambault, Beauchesne, & Caruso, 2011). If a journal was not included in the classification, the subfield most frequently occurring in its cited references was substituted. This is used as a control variable to control for differences in overall probability of recommendations across disciplines and subfields
- the average age of cited references in years
- the average citation count of the cited references
- the number of self-citations and non-self-citations 5 years after publication. The 5-year period was calculated with month-precision starting from the month of publication, as otherwise citation counts will be biased in favor of papers published early in a year and against those published late in a year (Donner, 2018)

### 3. Results

The dataset studied here comprises 2,122,550 publication records, of which 39,272, or 1.85 % have one or more F1000Prime recommendations. The papers are from 94 subfield classes of the SMJC, the most frequent of which are ,Developmental Biology‘, ,Neurology & Neurosurgery‘, and ,Oncology & Carcinogenesis‘. The overall average review duration (not journal-standardized) in the sample was 126 d (SD=92 d) and the average publication delay was 35 d (SD=45 d). For further analysis the two duration variables were transformed by calculating the standardized score (z-score) of each value of each paper in comparison only to all other papers in the same journal. This is justified as journals have widely diverging manuscript handling procedures and average turnaround times. A review duration of 2 months might be unusually long for one journal but unusually quick for a different journal. Journal-standardized publication delay and review duration were insubstantially correlated with  $r=0.03$ .

Table 1 shows the average journal-standardized scores, in standard deviation units, of different quality classes of papers according to the highest attained F1000Prime rating. For review duration, publications with any F1000Prime rating were reviewed around 2–3 % of a journal standard deviation faster than unrated items. For publication delay, the results suggest that only the very small class of papers rated “excellent” enjoyed some prioritization, as their time from acceptance to publication is about 3 % of a journal SD shorter than other papers’.

Table 1: Average journal-standardized review duration and publication delay by F1000Prime highest rating

rating	review duration	publication delay	observations
unrated	0.0005	0.0001	2,083,278
good	−0.0354	0.0019	18,554
very good	−0.0229	−0.0063	15,930
excellent	−0.0253	−0.0329	4788

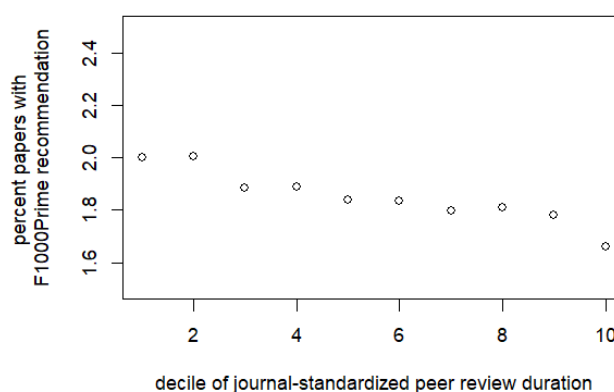
However, this analysis might be affected by quality stratification, as many journals have not published items from all four quality classes. Of the 1541 journals in the data set, only 305 have items of all three higher rating classes. To more appropriately study within-journal variation, Table 2 shows data analogous to Table 1 for the subset of these 305 journals. The results remain practically unchanged to the above ones. In either case, the observed differences are very small.

Table 2: Average journal-standardized review duration and publication delay by F1000Prime highest rating - subsample of 305 journals with items of all 3 higher ratings

rating	review duration	publication delay	observations
unrated	0.0012	0.0002	704,421
good	-0.0370	0.0048	13,553
very good	-0.0186	-0.0046	14,081
excellent	-0.0243	-0.0345	4734

We noted in the introduction that the functional form for the relationship between judged quality and review duration might conceivably be curvilinear as both high and low quality manuscripts might take longer to get reviewed than intermediate quality papers. To test this, journal-standardized peer review duration was binned into 10 intervals of equal size and for each bin the probability was calculated that a paper received any F1000Prime recommendation. This is shown in Fig. 1. The figure shows that the two decile groups with the shortest review duration are recommended at rates of about 2.0 % while the decile group with longest review durations has a recommendation rate of 1.66 %. The recommendation rates for the deciles in between are relatively smoothly declining from shorter to longer review durations. There is no evidence for a curvilinear relationship.

Figure 1: Percent of papers with F1000Prime recommendation across deciles of journal-standardized peer review duration



We now proceed to an analysis of predicting whether a paper has received any recommendation on F1000Prime from variables available primarily at publication time, including peer review duration and publication delay. We simplify the F1000Prime scores by considering a paper with any rating received as ‘rated’ and those without any as ‘unrated’. With this transformation, we can use standard logistic regression with dichotomous recommendation status as the dependent variable. Several dependent variables were transformed by taking logarithms and calculating standardized scores (z-scores) at the journal or global level, see below. We use the Tjur (2009) coefficient of discrimination to assess explained variation, a specialized quasi- $R^2$  value for logistic models.

A series of five models with different specifications were estimated. They are reported in Table 3 (as are applied variable transformations). Model 1 is a baseline model which includes only the control for subfield, as do all models, and the average JIF of the journal. Recommendation status can be predicted moderately well by these two variables alone ( $R^2=.14$ ). Trying to predict recommendation status from journal-standardized review duration and publication delay alone is not successful (model 2). The coefficient for publication delay is not significant at a 0.01 probability level. That for peer review duration is statistically significant but the proportion of explained variation is insubstantial ( $R^2=.02$ ). Model 3 combines average JIF, review duration and publication delay to test if the latter two variables provide any incremental explanatory power over the JIF-only model 1. This is not the case, the  $R^2$  remains nearly unchanged. Model 4 includes various other variables which are available at time of publication to test their incremental contribution to predicting recommendation status. The three cited-reference predictors are statistically significant, number of authors and number of countries, collaboration variables, are not. They are able to explain only a very small part of additional variation.

Model 5 represents a kind of best-case scenario as it includes in addition to all earlier discussed variables also the self-citation counts and non-self-citation counts after 5 years. Both variables are significant at the 0.01 level and their inclusion has several effects on other variables' coefficients: it renders review duration, number of cited references and average reference age insignificant, decreases the effect of average JIF, increases the effect of average citation count of cited references, and makes the coefficients for number of authors and countries become statistically significant. The size of the coefficient for non-self-citations is much greater than that for self-citations. Explained variation is increased only very slightly to  $R^2=.16$ .

In summary, there is a relatively strong relationship between average JIF and recommendation status, while other variables contribute little to nothing in incremental predictive value. The inclusion of individual paper citation counts, however, moderates the JIF-recommendation association.<sup>1</sup>

---

<sup>1</sup> As an aside, a model with only the two paper-level citation counts and without JIF has an  $R^2$  of .115 (not reported in Table 3). This indicates that while paper citation counts moderate and improve predictions of quality of JIFs, they do not by themselves provide for as much explained variation as JIFs for data from many journals.

Table 3: Results from logistic regression models

	model				
	1	2	3	4	5
avg. JIF	0.139* (0.001)		0.139* (0.001)	0.138* (0.001)	0.081* (0.001)
publication delay (js)		−0.007 (0.006)	−0.009 (0.007)	−0.004 (0.007)	0.004 (0.007)
review duration (js)		−0.079* (0.009)	−0.097* (0.009)	−0.083* (0.010)	−0.019 (0.010)
5-year self-citations (l, s)					0.092* (0.007)
5-year non-self- citations (l, s)					0.809* (0.007)
nr. of countries (s)				−0.007 (0.004)	−0.034* (0.005)
nr. of authors (s)				0.003 (0.003)	−0.027* (0.005)
nr. of cited references (s)				0.131* (0.005)	−0.004 (0.005)
avg. reference age (s)				−0.255* (0.008)	0.007 (0.007)
avg. reference citation impact (s)				−0.059* (0.008)	−0.165* (0.010)
Observations	2,121,208	2,122,523	2,121,181	2,120,000	2,120,000
Log Likelihood	−150,772	−177,522	−150,712	−149,677	−141,256
Akaike Inf. Crit.	301,734	355,237	301,618	299,558	282,721
Tjur R <sup>2</sup>	0.142	0.019	0.142	0.145	0.166

Notes:

\*  $p < 0.01$ .

All models include controls for subfields, coefficients not shown.

j.s.: standard scores at the journal level

s.: global standard scores

l.: log-transformed

#### 4. Discussion and conclusion

Using a large dataset from the health and life sciences we have investigated the hypothesis that papers which are of such high quality to be recommended on a dedicated paper recommendation platform, approximately the best 1 to 2 % of a cohort, are reviewed quicker and are published quicker after acceptance. While we have found statistically detectable effects for the first variable, the effect size was found to be so small as to be not helpful in distinguishing very high from ordinary-to-low quality papers. Better papers don't get reviewed (substantially) faster or slower.

#### Open science practices

This paper uses proprietary data which cannot be openly shared or accessed.

## Acknowledgments

I thank F1000Prime and Stephan Stahlschmidt for providing the recommendations data set.

## Competing interests

The authors declares no competing interests.

## Funding information

No grant funding was used for this research.

## References

- Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PloS ONE*, 4(6), e5910.
- Aksnes, D. W., Piro, F. N., & Fossum, L. W. (2023). Citation metrics covary with researchers' assessments of the quality of their works. *Quantitative Science Studies*, 4(1), 105-126.
- Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics*, 66-77.
- Beck, J., Sandbulte, J., Neupane, B., & Carroll, J. M. (2018). A study of citation motivations in HCI research. *SocArXiv preprint*, <https://doi.org/10.31235/osf.io/me8zd>
- Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. *Journal of Informetrics*, 9(3), 419-429.
- Bornmann, L., Schier, H., Marx, W., & Daniel, H. D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, 6(1), 11-18.
- Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37(1), 34-36.
- Donner, P. (2018). Effect of publication month on citation impact. *Journal of Informetrics*, 12(1), 330-343.
- Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics*, 76(1), 169-185.
- Hilmer, C. E., & Lusk, J. L. (2009). Determinants of citations to the agricultural and applied economics association journals. *Applied Economic Perspectives and Policy*, 31(4), 677-694.
- Kousha, K., & Thelwall, M. (2022). Covid-19 refereeing duration and impact in major medical journals. *Quantitative Science Studies*, 3(1), 1-17.
- 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.

Rigby, J., Cox, D., & Julian, K. (2018). Journal peer review: a bar or bridge? An analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics*, 114(3), 1087-1105.

Shen, S., Rousseau, R., Wang, D., Zhu, D., Liu, H., & Liu, R. (2015). Editorial delay and its relation to subsequent citations: the journals Nature, Science and Cell. *Scientometrics*, 105, 1867-1873.

Tjur, T. (2009). Coefficients of determination in logistic regression models - A new proposal: The coefficient of discrimination. *The American Statistician*, 63(4), 366-372.

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

Zhou, H., Han, R., Zhong, J., Qin, C., & Zhang, C. (2023). Which Factors Have a greater Impact on Review Time? — A Case Study of Nature Communication. *Proceedings of ISSI 2023 – the 19th International Conference of the International Society for Scientometrics and Informetrics*, 3, 111–112. <https://doi.org/10.5281/zenodo.10651692>