Does the Market of Citations Reward Reproducible Work?

Edward Raff
Raff_Edward@bah.com
Booz Allen Hamilton
Columbia, Maryland, USA
University of Maryland, Baltimore County
Catonsville, Maryland, USA

ABSTRACT
The field of bibliometrics, studying citations and behavior, is critical to the discussion of reproducibility. Citations are one of the primary incentive and reward systems for academic work, and so we desire to know if this incentive rewards reproducible work. Yet to the best of our knowledge, only one work has attempted to look at this combined space, concluding that non-reproducible work is more highly cited. We show that answering this question is more challenging than first proposed, and subtle issues can inhibit a robust conclusion. To make inferences with more robust behavior, we propose a hierarchical Bayesian model that incorporates the citation rate over time, rather than the total number of citations after a fixed amount of time. In doing so we show that, under current evidence the answer is more likely that certain fields of study such as Medicine and Machine Learning (ML) do reward reproducible works with more citations, but other fields appear to have no relationship. Further, we find that making code available and thoroughly referencing prior works appear to also positively correlate with increased citations. Our code can be found at the following link: https://github.com/EdwardRaff/ReproducibleCitations/

ACM Reference Format:
Edward Raff. 2023. Does the Market of Citations Reward Reproducible Work?. In 2023 ACM Conference on Reproducibility and Replicability (ACM REP ’23), June 27–29, 2023, Santa Cruz, CA, USA. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3589806.3600041

1 INTRODUCTION
A reproducibility crisis has been called for many scientific domains, including artificial intelligence and machine learning [1, 12, 20, 41]. It is paramount that all disciplines work to remedy this situation and push for reproducible work both as good science, and to mitigate such crises. Such work has begun in various fields with different strategies [10, 16, 18, 26, 33], yet the incentive structure around reproducing is long, with conflicting terminology across fields and years [25], we will use both terms interchangeably as our study focuses exclusively on cases where a different team independently performs the same experiments to obtain the same/similar results.

Citations are the primary reward for academic outputs, and to our knowledge, only the work of [34] has ever considered studying the relationship between papers that reproduce and the number of citations received. They used data on replication results from the fields of Psychology [10], Economics [6], and Social Sciences [7]. Distressingly, they conclude that non-reproducing work is cited more than reproducing works.

Our work revisits this hypothesis and data, and draws a different conclusion. We will show in section 3 that there are methodological issues that prevent a robust conclusion from being formed with the data and approach presented in [34]. Next, we will propose a Bayesian hierarchical model to alleviate these issues and allow further insight into the citation/replication question by incorporating a model of the citation rate changing over time in section 4. In section 5 we show our model is a significantly better fit to the data, and concludes that citation rate is unrelated or positively correlated with reproduction success, depending on the field being studied. Finally, we will conclude in section 6.

2 RELATED WORK
The study of paper citation has a long and multi-disciplinary history [23, 27–29, 32, 35], with many works proposing different power law variants to describe the distribution of citations. Most work that has looked at citations over time are looking at population level changes in citation distributions [2, 40, 42]. We are aware of only one prior work that looked at the citation rate by year through studying the impact of publication-vs-arXiv [39]. This work also modeled citation rates as a Poisson, similar to [34], which we will argue is an inappropriate model for citation count data.

Used by [34] were negative citations, a type of citation classification that can provide further insight into behaviors and results. The taxonomy of citation types, their labeling, and prediction [22] are another lens through which insight may be gained, but is beyond the scope of our study.

Dietz et al. [11] produced one of the first applications of Bayesian modeling to the study of citation behavior and influences. Our task is different, and so our model bares little resemblance, but the overall strategy we argue is worth further study. Several latent factors exist in bibliometric study to which modern machine learning may yield benefits, and the scale of bibliometric data provides fertile ground to new and technical challenges to advance the field.

3 ISSUES WITH EXISTING MODELING
While the Negative Binomial model has been previously identified to empirically have better performance at citation prediction [38],
Figure 1: Plate diagram of our proposed citation-replicated model. The observations are done against a Negative-Binomial model, which allows the model to predict integer outputs without the limiting assumption of the Poisson model that the mean is equal to the variance. Each “Field” place is a parameter adapted to different fields of study (e.g., Medicine or Machine Learning). “Citation Styles” are shared across the population of papers, allowing the model to learn a set of “types” of citation rates (e.g., and instant hit that is highly cited quickly, a slow burn that steadily accumulates, or even a flat line that is not cited). The “Observations” plate incorporates the impact of being reproducible (or not) and unobserved citations due to future years. The priors allow partial pooling across the groups. See Algorithm 1 for more details.

the Poisson model is still very popular. We note though that there is an easier way to show the Poisson model is in fact, inappropriate, for the bibliometric research it is used. The Poisson model assumes the mean and variance are equal, and if the variance is larger than the mean, the model suffers from overdispersion that prevents meaningful results. A statistical test [8] confirms with $p < 0.001$ that this is the case for citation data, which in the data from [34] has a mean of 438 citations but a variance of 504,639.

While Serra-Garcia and Gneezy [34] used the Poisson model in their work on the connection between replication and reproducibility, we note there are additional factors that lead us to challenge their initial conclusion. First is the question, what is the goal we are trying to measure? [34] explored citation rate as a function of reproducibility, and the predictions of citation markets. The latter is a subjective evaluation of a paper by (knowledgeable) individuals and not a factor of the paper itself. For this reason, they down-selected from $N = 139$ instances down to just $N = 80$ instances that have the citation market predictions. We argue that this is conflating variables of varying type, and unwise for answering questions about citation rates as it sacrifices 42% of the data that could be used for citation rate inferences, but do not have market prediction experiments. If we limit our experiments to the same 80 instances, we observe results more similar to what was reported but believe that using all available data regarding citation rate to make conclusions is the better choice.

Our interest is solely in the nature of the reproducibility of a paper correlating with citation rate, for which we argue it is best to include all data — and we do not wish to intermingle the analysis with a factual variable (successful replication). Further, we add additional data that is available from the Medical domain to broaden the types of fields covered and the amount of data available to make a robust conclusion. We carry out the modeling approach of [34] with simple linear models of the cumulative citation rate to show modeling issues with the approach, after which we will introduce our Bayesian model.

| Reproduced | coef | p     |
|------------|------|-------|
| Poisson-GS | 0.0172 | 0.129 |
| Poisson-SC | 0.1138 | <0.001 |
| Poisson-SC+M | 0.5775 | <0.001 |
| NB-GS      | 0.0172 | 0.150 |
| NB-SC      | 4.4592 | <0.001 |
| NB-SC+M    | 0.5777 | 0.004 |

To demonstrate the lack of robustness to the prior methodology, we will perform several repetitions of the overall approach choosing between:

1. Using the Poisson model vs a Negative-Binomial model
Does the Market of Citations Reward Reproducible Work?

ACM REP ’23, June 27–29, 2023, Santa Cruz, CA, USA

2

Figure 2: Correlation between Google Scholar and Semantic Scholar in the number of citations for each document per year. After multiple-test correction all years were significantly correlated with $p < 0.001$ in all cases.

(2) Using the original Google Scholar (GS) citation count data provided vs citation data from Semantic Scholar (SC)

(3) Using the original data with (SC) additionally with reproduction results from the Medical domain, adding a fourth field (+M).

This provides six total results, presented in Table 1. With all available data in use one can see that in no case do we observe a negative indication that papers that fail to replicate are cited more. However, we do see inconsistent conclusions about the impact of replication itself. When using Google Scholar the conclusion is there is no relationship, and when using Semantic Scholar the conclusion is a strong relationship. This challenge is not a factor of these citation sources having a dramatic disagreement, as can be seen in Figure 2 both are highly correlated in the per-year citations of the documents. This issue is instead that of model fit, as the highest adjusted $R^2$ fit amongst the Negative Binomial models is 0.0039.

The source of this discrepancy is the inappropriate merging of all data sources into one pool. The papers selected from Economics, Psychology, Social Science, and Medicine were all selected with biases toward higher citation rates — largely through the selection of high-impact factor sources. The citation rate per field, or journal, is not the same, as shown in Figure 3. Imbalances in the number of papers from each source that happened to replicate or not amplify spurious noise, resulting in low model fit and unstable conclusions.

4 METHODOLOGY

To address these problems, we propose a Bayesian hierarchical model that incorporates the citation rate over time, rather than the cumulative total number of citations. Our interest in citation rate over time is of interest not merely for model fit, but primarily because we are interested in observing whether the types of citation patterns vary between reproducible and non-reproducible papers.

That is to say, some papers do not start to accumulate citations for a considerable amount of time, others reach a steady state of citations, and others reach a peak citation rate before their citation rate drops. A total-citation rate model can not reveal anything about this question.

The high-level plate diagram of our approach is presented in Figure 1, which we will discuss at a high level with the detailed generative story given by Algorithm 1. The coefficients $\beta$ are with respect to each Field, with a hierarchical prior used over them and a shared ridge regression penalty (variance of the Gaussian distribution).

The observations are done with respect to a zero-inflated Negative-Binomial model, parameterized with a mean and dispersion factor $\mu$ and $\varphi$ as shown in Equation 1. The zero-inflation serves two purposes. First, some papers do receive zero citations for some time before becoming popular, and the zero-inflation model prevents down-weighting the citation rate $\mu$ from these zero citations. Second, it allows us a convenient way to handle the fact that papers were published at different times, and thus for a desired horizon of

\[
\text{NegBinomial2}(n | \mu, \varphi) = \left( \frac{n + \varphi - 1}{\varphi} \right) \left( \frac{\mu}{\mu + \varphi} \right)^n \left( \frac{\varphi}{\mu + \varphi} \right)^\varphi. \tag{1}
\]
T years, not all papers will have T years of existence to accumulate citations. When a year has not yet occurred, we force the zero inflation gate to effectively mask the year with no impact on the model. We used a target of T = 10 years in all cases. Each paper receives its own gate value with a hyperprior shared over all samples. We use the proportional Beta hyper prior as shown in Equation 2 with a non-informative prior over μ.

\[
\text{BetaProportion } (\theta | \mu, \kappa) = \frac{1}{B(\mu \kappa, (1-\mu) \kappa)} \theta^{\mu \kappa - 1} (1-\theta)^{(1-\mu) \kappa - 1}
\]

(2)

To represent the impact of the t’th year’s citation rate of the i’th sample μ_{i,t} we model a base citation rate μ_i modulated by an annual base citation multiplier sampled from a Gamma prior centered at a mean of 1.0 (i.e., no change in annual citation rate). The impact of the compounding base rate can be delayed (but not increased, as that implies pre-publication citations) by a shift factor samples from a positive Laplacian scaled so that the entire T years may be selected by the prior would prefer no shift.

We do not give each sample its own base and shift as it allows significant over-fitting of the model to ignore the impact of the coefficients β. Instead, we use a Dirichlet process to sample from a pool base/shift pairs — where reproducible and non-reproducible papers each receive a separate Dirichlet process sampling from the same pool. We enforce a sparse process by putting a Beta prior over the α parameter of the processes so that we may see if there is a difference in the types of citation styles between papers (e.g., do non-replicating papers more frequently have decaying base rates < 1?).

The full model is detailed in Algorithm 1. We use NumPyro [24] to implement the model with the NUTS sampler [19]. In all cases, we use 500 burn-in iterations followed by 2,250 steps with a thinning factor of 3.

5 RESULTS

Now that we have specified our approach to understanding how citations may be impacted by a paper’s ability to replicate, we will present our results in two sections. First, we will consider the results with respect to the previous fields of study (Medicine, Economics, Psychology, and Social) and show that we obtain consistent results and reasonably believe them to be a more reliable model. Second, we will repeat the study applied to data from machine learning [30]. This data is studied separately because it has a different kind of selection bias, and a different set of available features to consider, then the other data.

5.1 Science Results

We begin by examining the conclusions inferred by our model on the three versions of the data. Google Scholar, Semantic Scholar, and Semantic Scholar with the medical domain added. The results can be found in Figure 4, showing consistent conclusions of no correlation between field and citation rate of reproducible papers for any of the three original fields. When Medicine is added we observe that it does show a high citation rate for reproducing papers, without changing the conclusion of the other fields.

Algorithm 1 Our Hierarchical Bayesian generative story for modeling citation rates. The + indicates distributions truncated to be non-negative.

\[
\text{Require: } N \text{ observations with } r_t \in \{S, F\} \text{ for successful or failed reproduction and } f_i \text{ indicating the field of research for the paper.}
\]

\[
\lambda_{ridge} \sim \text{HalfCauchy}(0,1) \quad \phi \sim \text{Beta}(1, 10) \quad \omega^S \sim \text{Dirichlet}(\alpha) \quad \omega^F \sim \text{Dirichlet}(\alpha)
\]

\[
\text{for all } i \in 1, \ldots, \infty \text{ do}
\]

\[
\text{Citation Style for } \omega^S \text{ will sample from }
\]

\[
\text{shift } \sim \text{Laplace}^+ (0, \text{years out/6})
\]

\[
\text{base } \sim \Gamma(100, 100) \quad \text{This Gamma distribution will encourage values near } 1, \text{ as values > 2 are undesirable in being unrealistic.}
\]

\[
\text{end for}
\]

\[
\text{Hierarchical Reproducible Prior}
\]

\[
\text{for all Field of Study } i \text{ do}
\]

\[
\phi_i^\text{field} \sim N(\phi_i^\text{field}, \lambda_{ridge}) \quad b_i \sim \text{Cauchy}(0, 1) \quad \text{Bias term is independent between Fields}
\]

\[
\text{end for}
\]

\[
\text{Uninformative prior on the mean rate of no citations occurring.}
\]

\[
\phi_i^\text{gate} \sim U(0, 1) \quad \text{Select the citation style base/shift for this sample based on the distribution w.r.t. the sample replicating or not}
\]

\[
\log(\mu_i) \sim \phi_i^\text{field} \cdot 1[r_i = S] + b_i \quad \text{The rate is modified based on the paper replicating or not.}
\]

\[
gate_{t} \sim \text{BetaProportion}(\text{gate}^\text{base}, \text{gate}^\text{shift})
\]

\[
\text{for all Time steps } t \text{ do}
\]

\[
\mu_i \sim \mu_i \cdot \text{base}^\text{max}(t - \text{shift}_t, 0)
\]

\[
\text{accumulate Zero-Inflated Negative Binomial loss}
\]

\[
\text{NetBinomial2}(y_t | \mu_i, \phi) \text{ with gate probability } \gate_t
\]

\[
\text{end for}
\]

\[
\text{end for}
\]

Beyond the consistency of the conclusions, we are further confident in our approach’s conclusions due to better model fit. The Google Scholar case produced an $R^2 = 0.41$, and the Semantic Scholar data with/without the Medicine papers at $R^2 = 0.24$ and 0.19 respectively. We arguably would not expect very high $R^2$ values considering the model is characterizing populations of citation rates based only on the field, as prior work focusing on predicting citations using the venue, author, and content information achieved $R^2 = 0.74$ [43].

This approach has also provided further insight into the nature of reproduction and citations, that the reward behaviors are not consistent across fields. The question then becomes: do reproducible papers have a different style of citation patterns (i.e., accumulating or decreasing citation rates at a different pace) compared to non reproducible work?
Figure 4: The results of the coefficients $\beta$ for the different fields of study when using Google Scholar data (top), Semantic Scholar (middle), and Semantic Scholar with the addition of the medical papers (bottom). The x-axis is the coefficient value and the forest plot shows the estimated value and 95% credible interval.

Per the design of our model, in Figure 5 we can investigate the citation rates over time as inferred by our model, shown for the Semantic Scholar + Medicine case. In this instance, we do not observe any difference in the citation rates or style between (non)reproducing papers. A maximum of 50 components were allowed for computational tractability, and non-present components are ones the model learned to discard with near-zero probabilities.

We note of particular interest that most latent citation styles only have an impact starting two years out from publication, a result consistent with prior work which found the first two years of citations to be highly predictive of the long-term cumulative number.
of citations [36]. This provides another degree of confidence in the validity of our general approach, though we do not make a claim that our simple model of citation rate is the best possible choice.

The data is also interesting in that we observe behaviors not normally discussed in bibliometric literature: papers whose citation rate decreases with time. This is indeed not directly observable in the common modeling approach of looking at cumulative citations after a point in time. We further find citation style 29 uniquely interesting as a "runaway success", quickly multiplying the citation rate by \( \exp(10) \approx 10^{4.35} \) after ten years.

### 5.1.1 Results when using Limited Data

As noted previously, [34] used a reduced set of the total datasets, selecting conditional on a betting market having been performed for the given paper. This is unrelated to whether or not the paper actually replicated, and so we do not believe it is an appropriate criterion for the analysis. However, if we repeat our approach with the same selection, we observe that some fields result in a negative correlation while others retain no relationship, and medicine remains positive. This can be observed in Figure 6.

![Figure 6: Results when removing 42% of the papers with replications in the manner performed by [34]. This shows how they obtained the negative correlation, but their selection criteria is not related to actual replication success or lack of citation data. For this reason, we consider it an unlikely influence.](image)

### 5.2 Machine Learning Results

Having shown our model allows for more robust conclusions around the impact replicable results have on citation rate, we turn the machine learning reproductions documented by [30]. Many of the papers were selected by the author’s personal interests, rather than impact factors, so we do not find it appropriate to include them in the same hierarchical model. The ML data also includes numerous other quantifications about the paper not present in the prior section, so we treat it separately. We use the same approach without a hierarchical prior since it is one population of papers. The adjusted \( R^2 \) of the model is 0.31 using Semantic Scholar for the citation data, in line with the prior experiments.

![Figure 7: Forest plot of the coefficients \( \beta \) of various features, with 95% credible intervals. The results notably show that in machine learning reproducible papers are cited more. In addition, making code available, and citing more prior work, intuitively correlate with more citations. Other variables such as the number of tables, use of conceptualization figures, and Journal vs Workshop publication have difficult to intuit correlations.](image)

The results Figure 7 show that reproducible papers, and papers that make their code available, both receive higher rates of citation. The former is desirable, and the latter indicates a strong motivation for authors to open source their code beyond the arguments around replication [5, 9, 16, 17, 21]. The sharing of code is generally argued to be beneficial, but we do note that it captures methodological flaws as well — and is thus not a panacea to concerns around reproduction [3, 4, 13–15, 37]. We are also encouraged that more references per page have a higher citation rate, under the belief this corresponds...
to more thorough documentation of prior work and good scholastic behaviors.

Figure 8: The discovered latent citation styles and their proportion of use in reproduced and failed to reproduce papers (top) and the log multiplicative effect of the citation rate over time (bottom) for the Machine Learning data. Note the bottom legend shows “Citation Style, Mean Occurrence Rate of No Reproduction”.

The reduced citation rate for Conceptualization Figures, which attempt to convey the intuition of a method, is interesting. Raff [30] noted no relationship between this variable and replication, while later work found that papers that use conceptualization figures take less time/human effort to reproduce [31]. This type of scientific communication appears to have a particularly complex relationship with reproduction and the incentives around reproduction that thus warrants further study.

The last points of note are that publishing in Journals, and more tables appear to increase citation rate while publishing in a workshop reduces it. Publishing in a workshop having a lower citation rate makes sense intuitively, though it is perhaps interesting that tech reports (like arXiv) have no relationship — and it is worth studying whether workshops being a final “home” for a paper may have a negative perception. This result is also possibly due to the noted bias in the data, which we believe may explain the result that Journal publications have a higher citation rate, since ML as a field generally prefers conferences over journals. Last, we have no particular intuition about why having more tables per paper may lead to more citations — unless it is a matter of making it easy for future papers to re-use the reported results, a hypothesis proposed in [30].

Last, we look at the latent citation patterns again in Figure 8, and note that style 29 does have a significant difference between reproducible and non-reproducible works1. By chance this component again represents a “runaway success”, indicating a preference for a degree of success toward reproducible works.

6 CONCLUSION

Our results are overall encouraging toward the question of replication and citation: citations do appear to increase or are independent of replication, which is better than the prior hypothesis that non-reproducible works get more citations. Our results for machine learning in particular indicate that citations correlate positively with further desirable behaviors like thorough citations and sharing of code. This work has furthered the bibliometric study of the interaction between citation rate and replication, and we note further valuable directions remain. A large amount of data without ground-truth replication success exists to merge into such analyses, as well as the possibility of using natural language processing to make inferences about paper replications by the content of citing documents.

REFERENCES

[1] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 7604 (2016), 452–454. https://doi.org/10.1038/533452a
[2] Lutz Bornmann and Rüdiger Mutz. 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 (2015), 2215–2222. https://doi.org/10.1002/asi.23329
[3] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyopruik, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Veleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. 2021. Accounting for Variance in Machine Learning Benchmarks. In Machine Learning and Systems (MLSys). http://arxiv.org/abs/2103.02098
[4] Xavier Bouthillier, Cézar Laurent, and Pascal Vincent. 2019. Unreproducible Research is Reproducible. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97). Kamalika

1This is a different style “29” than in the prior section, it is by pure chance that it also happened to be the 29th component of the same nature. This took us many hours to "debug" to find no apparent bug, just random chance.
[21] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussmann, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupiter development team. 2016. Jupiter Notebooks - a publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas, Fernando Louâidès and Burgit Schmutz (eds.) IOS Press, 87-90. https://eprints.soton.ac.uk/409313/

[22] Suchetha K Nunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. 2022. A meta-analysis of semantic classification of citations. Quantitative Science Studies 2, 4 (2 (2022)), 1170–1215. https://doi.org/10.1162/qss_a_00159

[23] Alfred J Lotka. 1926. The frequency distribution of scientific productivity. Journal of the American Society for Information Science 16, 12 (10 1926), 317–323. http://www.jstor.org/stable/2429203

[24] Du Phan, Neeraj Pradhan, and Martin Jankowiak. 2019. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. arXiv (2019), 1–10. http://arxiv.org/abs/1912.11554

[25] Hans E Fissler. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Frontiers in neuroinformatics 11 (1 2018), 76. https://doi.org/10.3389/fninf.2017.00076

[26] Russell A Pollack. 2019. The Costs of Reproducibility: Neuron 101, 1 (1 2019), 11–14. https://doi.org/10.1016/j.neuron.2018.11.030

[27] William Gray Potter. 1981. Lotka’s Law Revisited. Journal of Research in Science Education 11, 4 (1 1981), 347–364. https://doi.org/10.1016/0304-4076(80)90014-K

[28] Derek De Solla Price. 1976. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27, 5 (1976), 292–306. https://doi.org/10.1002/asi.4630270505

[29] Derek J de Solla Price. 1965. Networks of Scientific Papers. Science 149, 3683 (7 1965), 510–515. https://doi.org/10.1126/science.149.3683.510

[30] Edward Rafe. 2019. A Step Toward Quantifying Independence in Reproducible Machine Learning Research. in NeurIPS. http://arxiv.org/abs/1909.06674

[31] Edward Raff. 2021. Research Reproducibility as a Survival Analysis. In The Thirty-Fifth AAAI Conference on Artificial Intelligence. http://arxiv.org/abs/2102.09932

[32] S Redner. 1998. How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B Condensed Matter and Complex Systems 4, 2 (1998), 131–134. https://doi.org/10.1007/s1005100500359

[33] D Scullley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Eberas, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS’15). MIT Press, Cambridge, MA, USA, 2503–2511.

[34] Marta Serra-Garcia and Uri Gneezy. 2021. Nonreproducible publications are cited more than replicable ones. Science Advances 7, 21 (5 2021), eabd1705. https://doi.org/10.1126/sciadv.abd1705

[35] William Shockley. 1957. On the Statistics of Individual Variations of Productivity in Research Laboratories. Proceedings of the IRE 45, 3 (1957), 279–290. https://doi.org/10.1109/JRPROC.1957.287864

[36] Clara Stegehuis, Nelly Latvak, and Ludo Waltman. 2015. Predicting the long-term citation impact of recent publications. Journal of Informetrics 9, 3 (2015), 642–657. https://doi.org/10.1016/j.joi.2015.06.005

[37] Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Gong. 2020. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In Fourteenth ACM Conference on Recommender Systems (RecSys’20). Association for Computing Machinery, New York, NY, USA, 21–23. https://doi.org/10.1145/3383313.3412489

[38] Mike Thelwall and Paul Wilson. 2014. Regression for citation data: An evaluation of different methods. Journal of Informetrics 8, 4 (10 2014), 963–971. https://doi.org/10.1016/j.joi.2014.09.011

[39] Y A Traag. 2021. Inhering the causal effect of journals on citations. Quantitative Science Studies 2, 2 (7 2021), 496–504. https://doi.org/10.1162/qss_a_00128

[40] Attila Varga. 2019. Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers. Proceedings of the National Academy of Sciences of the United States of America 116, 44 (2019), 22094–22099. https://doi.org/10.1073/pnas.1905819116

[41] Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. Voodoo Correlations in Social Neuroscience. Perspectives on Psychological Science (2008).

[42] Matthew L Wallace, Vincent Larivière, and Yves Gingras. 2009. Modeling a citation advantage processes. Journal of the American Society for Information Science 59, 3 (2008), 302–309. https://doi.org/10.1002/asi.20841

[43] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation Count Prediction: Learning to Estimate Future Citations for Literature. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM ’11). Association for Computing Machinery, New York, NY, USA, 1247–1252. https://doi.org/10.1145/2063576.2063577