Misapplied Metrics:
Variation in the *h*-index within and between disciplines

Misapplied Metrics:
Variation in the *h*-index within and between disciplines

Abstract

Scholars and university administrators have a vested interest in building equitable valuation systems of academic work for both practical (e.g., resource distribution) and more lofty purposes (e.g., what constitutes "good" research). Well-established inequalities in science pose a difficult challenge to those interested in constructing a parsimonious and fair method for valuation as stratification occurs within academic disciplines, but also between them. Despite warnings against the practice, the popular *h*-index has been formally used as one such metric of valuation. In this article, we use the case of the *h*-index to examine how within and between discipline inequalities extend from the reliance of metrics, an illustration of the risk involved in the so-called "tyranny of metrics." Using data from over 42,000 high performing scientists across 120 disciplines, we construct multilevel models predicting the *h*-index. Results suggest significant within-discipline variation in several forms, including a female penalty, as well as significant between discipline variation. Conclusions include recommendations to avoid using the *h*-index or similar metrics for valuation purposes.

Keywords

Metrics; *H*-index; Gender Inequality; Team Science; Higher Education

Introduction

From teaching to research, systematic evaluation of academic work presents a unique set of challenges as the academic disciplines that broadly organize scholarly labor differ upon several relevant factors, including demographic characteristics and the prevalence of team science for example. Differences in demographic characteristics, like race and gender, capture processes of racism and sexism that limit the opportunities for some scholars and not others across disciplines (Hofstra et al., 2020; Larivière et al., 2013; Xie & Shauman, 1998). Cultural and economic differences between disciplines, like the prevalence of team science or even the definition of success, can affect the quality and quantity of publication (Fortunato et al., 2018; Gardner, 2009; Stephan, 2012). Scholars have invented dozens of metrics that are formally or informally used in the evaluation of scholarly research, despite warnings about apples to oranges comparisons given within discipline and between discipline differences (Hicks et al., 2015). The most widely used metric of this type is the *h*-index, a simple score where a scholar with 10 published articles with at least 10 citations receives a *h*-index of 10.[1] The limitations of the *h*-index are well understood, but less is known about how processes of inequality are embedded within its distribution within and between disciplines. Here, we ask two questions: 1.) What within discipline factors contribute to variation in the *h*-index? 2.) Do between discipline differences exist in the *h*-index?

These questions arise from at least two important considerations. First, a defining characteristic of academic labor is that, like other forms of work, it is beset by inequalities. A large body of research describes the factors that contribute to inequality in academia (see Fox et al., 2017; Long & Fox, 1995 for reviews). From well-known pipeline effects to publication differences to varying effects of parenthood, scholars have shown the multifaceted ways some scholars are obstructed from paths to success at work (e.g., Fox, 2005; Grant et al., 2000; Morgan et al., 2021 among many others). At the same time, differences exist in the distribution of resources to disciplines as can be seen in variation in federal funding and the way resources are distributed to departments, the proxies for disciplines on university campuses (Katz & Matter, 2020). Second, numerous scholars have drawn attention to the risks of the increasing push to quantify aspects of social life, the so-called "tyranny of metrics" (Muller, 2019). This research points to the crude ways that quantification reduces complex social phenomena often with implications for inequality and nearly always with implications for valuation or what is considered good or bad. In the case of scholarly metrics, critics have raised concerns about how these metrics transform scholarship into a capitalist-like market (e.g., neoliberalization), build unproductive individual constraints into academic labor, such as the anxiety connected to hyper-competitiveness, and ultimately result in less productive and innovative scholarship (Edwards & Roy, 2017; Muller, 2019) and efforts to "game metrics" at either the journal or individual-level (Siler and Larivière,

[1] We reluctantly use the "*h*-index" label throughout for the metric introduced by Hirsch (2005) while acknowledging this anonymization abstracts the sociohistorical context in which it was created. We occasionally offer the Hirsch Index to maintain Jorge Hirsch's role in its creation.

2022). Turning to within and between discipline variation in the *h*-index offers insight into how metrics relate to inequalities and provide a case of the potential for tyrannical metrics.

To understand factors that contribute to differences in scholar's *h*-index or Hirsch Index, we turn to data on over 42,000 high-performing scientists across 120 disciplines. This analysis draws from Ioannidis et al.'s (2019) dataset identifying the top scientists in the Scopus database. To understand within and between discipline differences, we construct multilevel models predicting the *h*-index with a range of factors including disciplinary age, the number of publications, the number of sole publications, a female name index, and measure of university productivity. This approach is consistent with calls to develop multilevel analyses of gender disparities in science (Fox, 2020). Results indicate that significant within discipline differences exist, including a female and sole authorship penalty. Results further show significant between discipline differences with a range of estimated mean *h*-index scores from 22 to 68. Conclusions include recommendations to avoid using the *h*-index or similar metrics for valuation purposes.

Inequality in Science

Inequality in science, and in academia generally, persists within and between disciplines and along well-known axes. Gender is one of the most studied axes of inequality in science. Research on the scientific pipeline, for example, shows how obstacles, like cultural stereotypes about skills differences in math and science and related social psychological factors such as a sense of belonging, limit pathways to particular scientific fields for girls (Cech et al., 2011; Ma & Xiao, 2021; Penner & Willer, 2019). When these significant obstacles are overcome, women continue to experience inequality in academic work, including in publication. Publication is both the outcome of academic labor and the currency of academic careers. How many and the type of publications produced by a scientist translates into tangible resources, like salary raises and the job security of tenure, and less tangible resources, like prestige. While significant gains have been made by women in academic work - numerous fields that were predominantly men in the mid- to late-twentieth century are now majority women - publication, both in the quantity and quality, remains a site of persistent inequality (Xie & Shauman, 1998).[2]

Sociologists of science have spent decades trying to disentangle the factors associated with gender differences in publication, especially related to differences in the number of publications and citations. Nearly 40 years ago, Cole and Zuckerman (1984) referred to the ongoing male advantage in publication and citation counts as the "productivity puzzle" as the causes of this advantage remained difficult to pinpoint. They conclude, "[S]ince gender differences in published productivity persist, the productivity puzzle has yet to be solved" (p. 250). While much has changed as women have made significant gains in many academic fields, research suggests that the productivity puzzle remains. For example, Leahey shows how the level of

[2] Early signs indicate that these forms of gender inequality may have increased due to the COVID-19 pandemic or a "pandemic penalty" (King & Frederickson, 2021).

research specialization intervenes with gender affecting productivity with consequences for earnings (Leahey, 2006, 2007). More recently, Hofstra et. al. (2020) provide evidence of the complex mechanisms that help perpetuate gender and racial hierarchies in academic work. Using data on US doctoral recipients, they find that gender and racial minorities are more likely to generate innovative scientific work, but that this work is less likely to be adopted by future researchers with consequences for academic hiring.

A prestige puzzle may also follow the productivity puzzle as women may be less likely to publish in the top or most prestigious journals in their fields. As top journals have higher impact factors, the prestige puzzle could have a significant impact on differences in publication metrics and career outcomes. This form of inequality may occur for a variety of reasons from a lack of mentorship, different family and work-based responsibilities between men and women, and differences in specialization similar to Leahey's work on specialization and productivity (Light, 2009). Drawing on literature on occupational segregation and identity, Light (2013) shows how the prestige puzzle has changed over time within sociology. Earlier cohorts of women sociologists were significantly less likely to publish in top sociology journals compared to men regardless of specialty areas. However, as more women entered sociology, occupational segregation occurred with subfields becoming sharply gender imbalanced. While baseline models of the prestige puzzle for more recent cohorts reveal the persistence of this form of inequality, more complete models that control for specialization, or occupational identity, show that the contemporary effect likely operates through these segregation processes.

Collaboration also likely plays a role with significant historical differences in coauthorship networks based on gender (Moody, 2004). More recent work identifies how team science affects gender and publication in both political science and sociology pointing to how the structure of scientific work can negatively impact women social scientists (Akbaritabar & Squazzoni, 2021; Teele & Thelen, 2017). This suggests how within discipline differences may affect publication metrics, like the *h*-index, but also points to how inequality may occur between disciplines as pronounced differences exist between disciplines in terms of factors like gender composition and team vs. sole authorship.

Inequality Between Disciplines

Both differences in gender composition and differences in the frequency of team science may have direct and indirect effects on how resources are distributed in universities. However, less is known about how these between discipline-based inequalities differ from within discipline inequality. Research on collaboration and citation impact using the *h*-index shows disciplinary differences in the effect that collaboration has on impact with more collaboration having a stronger impact in physics and medicine, while having a smaller effect in the brain sciences or computer science (Parish et al., 2018). Disciplinary differences occur regarding more tangible resources, like federal funding. Lanahan et al. (2016) describe the substantial differences between fields in terms of federal funding with implications - a "domino effect" - for future non-

federal funding and a potential site of cumulative advantage stratifying disciplines. One of the ways that disciplines differ and also one of the ways that disciplines may be valued differently is how they perform on commonly used metrics.

Tyranny of Metrics

The tyranny of metrics is the process by which quantification becomes the central factor in determining worth. This "metrics fixation" occurs in a variety of fields from the law to business to education (Muller, 2019, p. 4). The turn to metrics is part of the broader process of neoliberalization of education. While neoliberalization is widely and sometimes sloppily used and therefore, perhaps, easy to dismiss, the concept succinctly captures the spread of economics logic - an effort to "economize everything" (Berg et al., 2016, p. 171) - such that neoliberal reason becomes common sense or simply the default rationale people use to make decisions. Metrics reinforce the notion that individual performance at work can be easily calculated and compared; therefore, material rewards like promotions and raises can be fairly and transparently applied. Of course, metrics often hide as much as they reveal as they simplify a process by carving away essential components. In universities, the metrics that help reinforce and are reinforced by neoliberal reason summarize entire careers for administrators who likely have little to no understanding of the research that the metrics summarize. This disempowers colleagues within and outside of a scholar's university, while centralizing power in the hands of bureaucrats with little incentive to actually understand academic work.

University administrators' use of metrics to evaluate faculty output is a fairly recent phenomenon. Prior to the development of bibliometric indicators, evaluation of scholarly research was performed primarily by disciplinary specialists who offered qualitative assessment of a research record. While peer assessment is still a central part of evaluation processes, metrics such as citation counts, journal impact factors, and the *h*-index are now commonly incorporated into hiring and promotion decisions (McKiernan et al., 2019) and are seen as more important to more vulnerable untenured scholars than tenured ones (Niles et al., 2020). Qualitative expert assessment that involves hours of engagement with the work summarized by metrics is likely seen with suspicion by some administrators motivated by a logic that privileges competition, central decision-making, and market valuation (Berg et al., 2016).

Critics have raised concerns about how metrics transform scholarship into a capitalist-like market at the core of neoliberalization resulting in "perverse incentives" for researchers to publish shoddy or fraudulent work (Edwards & Roy, 2017), while simultaneously resulting in mental health trauma for academic workers experiencing hypercompetitive markets and suspicious management (Forrester, 2021). Administrators' increasing reliance upon metrics serves as a modern form of Taylorism, a production principle used in the early twentieth century that pursued technological solutions to the "problem" of worker-related inefficiencies on the shop floor with little regard to employee satisfaction or wellness. By using technology to set the

nature and pace of production, owners and managers gained greater control over the labor process itself. Prioritizing metrics creates a demand for quantity over quality, and by following these demands academic laborers risk surrendering some degree of control over their own labor processes.

The Case of the h-index

One key metric used for evaluation purposes is the *h*-index or Hirsch Index. Physicist Jorge Hirsch proposed the *h*-index as a "useful index to characterize the scientific output of a researcher" in a 2005 article in the *Proceedings of the National Academy of Sciences* (Hirsch, 2005). While acknowledging the "potentially distasteful" use of metrics for evaluation, he presents quantification as an economical means of evaluating impact. In this highly cited article, Hirsch defines the *h*-index as follows: "A scientist has index *h* if *h* of his or her N_p papers have at least *h* citations each and the other ($N_p - h$) papers have $\leq h$ citations each" (p. 16569). In other words, a scholar with 10 of their 100 publications with a citation count of 10 will have a *h*-index of 10. He goes on to specify - again with some acknowledgement that metrics offer a "rough approximation" of a research portfolio - how and when the index could be put to use: "Based on typical h and m values found, I suggest (with large error bars) that for faculty at major research universities, $h \approx 12$ might be a typical value for advancement to tenure (associate professor) and that $h \approx 18$ might be a typical value for advancement to full professor" (p. 16571). In sum, this publication announced a simple means of evaluating research impact and permission to use the metric for evaluation purposes.

The immediate response to the *h*-index was largely positive with features in top scientific journals; however, some criticism of the index also quickly appeared (Barnes, 2017). Critics identified a range of issues from the relationship between the *h*-index and career length as well as the effect of self-citation (see Kelly & Jennions, 2006; Purvis, 2006 among numerous others). However, the *h*-index and variants have proven enormously popular both in the bibliometrics and science of science communities and among university administrators seeking quick and cheap ways to evaluate scholars, including universities and science funding agencies (Barnes, 2017). The *h*-index is included as a key quantitative metric for annual review and/or tenure and promotion in faculty handbooks in a range of departments and schools in the United States (c.f., handbooks from the Boston University School of Public Health (Boston University, 2018), the Ohio State University Department of Surgery (Ohio State University, 2014), or Oregon State University's College of Business (Oregon State University, 2020)). Survey research in Germany on whether and how scholars understand the importance of the *h*-index indicate that scientists widely understand the importance of the *h*-index to their careers, but scholars in the humanities and social sciences do not (Kamrani et al., 2021). This variation is unfortunate as quantitative metrics are widely applied in German universities as elsewhere. In sum, despite some criticism, the *h*-index has been widely adopted although perhaps not widely understood. This analysis intends to examine within and between discipline inequalities in the *h*-index.

Hypotheses

We develop the following hypotheses to better understand within and between discipline differences in the *h*-index, or Hirsch Index, based on the prior literature. Consistent with work on gender inequality and science and particularly work on the productivity and prestige puzzles, we examine the following:

> Female Penalty Hypothesis (H1): Authors with names more frequently associated with women are more likely to have *lower h*-index scores.

Inlight of research on team science and its impact on academic careers, we examine the following:

> Sole Author Hypothesis (H2): Authors with more sole-authored publications are more likely to have *lower h*-index scores.

Last, although less well understood, based on the research describing disciplinary differences in culture, such as propensity to collaborate, and material differences in resources, we examine the following:

> Disciplinary Differences Hypothesis (H3): Significant *h*-index differences likely exist between disciplines.

Data and Methods

To evaluate these hypotheses, we analyze data on high performing scholars according to well-known metrics. Ioannidis et al. (2019) collected author-level bibliometrics on 100,000 high performing scholars from the Scopus database. They updated this data through 2019 and expanded to include the top 2% of authors in a wide range of disciplines (Ioannidis et al., 2020). We use the most up-to-date datafile for these analyses. These data suffer from several limitations, importantly including database coverage with some likely disciplinary differences (Mongeon & Paul-Hus, 2016; Singh et al., 2021). Nonetheless, these well-curated data represent a unique opportunity to evaluate within and between disciplinary differences. We also see these data as offering a conservative test of these differences as variation is likely to widen when moving beyond scholars who are in the top 100,000 on these metrics.

We reduce the data along several dimensions to address some of the limitations of the Scopus database and for analytic purposes. We limit the data to those who first published in 1960 or after and those who last published in 2015 or before to reduce potential noise in the data and provide an adequate picture of a plausible late to early stage academic career. We also limit the data to authors in the United States to account for geographic variation in the Scopus database. We also filter the data to exclude the Visual & Performing Arts, Philosophy & Theology, Built Environment & Design, Historical Studies, Communication and Textual Studies fields due to their relative rarity and to reduce apples to oranges comparisons. Disciplines were also filtered

for size (n>100) on top of the field filters. These reductions, plus listwise deletion of omitted variables, results in a sample size of 42,509 scholars nested in 120 disciplines.

The focus of the analysis is the *h*-index, the dependent variable. Independent variables include: percentage female name, percentage of sole publications, number of citations from 1996-2019, the university count of citations allstars, the disciplinarity score, and publishing age or the years between first and last publication. Percentage female name is constructed using the gender package in R (Blevins & Mullen, 2015). Estimating gender is problematic for numerous reasons including the typical reliance on government-provided data that often assumes and contributes to the gender binary (see Mihaljević et al., 2019), and these methods should be used with caution and only when necessary. In this case, names are our only potential means of estimating gender. Here, we use the "ssa" method in the gender package drawing on data first names from 1940-1990 from the US Social Security Administration. As an acknowledgment of the imperfect assignment of gender using these methods, we use the proportion of female names for this range of years and scale by 100 for the percentage female name, rather than an arbitrary assignment of binary gender. The percentage of sole authored publications is constructed by dividing an author's number of sole authored publications by their total publications and we multiply this proportion by 100. We scaled the number of citations by dividing by 100. The university count is the number of scholars from each author's university in the complete dataset. The disciplinarity score is the proportion of an author's articles that are situated in their most frequent field. We scaled this proportion by multiplying by 100.

To further evaluate the within and between discipline factors related to the *h*-index given the limitations of the Scopus database, we run comparable analyses over a subset of the top-performing scholars who specialize in the field of clinical medicine. The Scopus database's coverage of this field is considered strong relative to the social sciences and humanities (Mongeon & Paul-Hus, 2016). This subset, filtered for disciplines with at least 25 scholars, has a sample size of 16,041 nested in 44 disciplines.

Table 1 presents the descriptive statistics for the variables used in both the full (1a) and clinical medicine (1b) analyses. Note several differences between the two datasets with scholars in the clinical medicine set having a higher *h*-index, higher number of citations, and lower percentage of sole author publications consistent with a more strictly natural or life science set, but a similar mean percent female name.

<Table 1 about here>

Analytic Strategy

These analyses use multilevel, or hierarchical linear, models as the data are nested: Each individual scholar is nested in a discipline. Multilevel models are more appropriate than fixed effects models in this case because they offer greater insight into the structure of variation or, here, on the relationship within and between discipline differences (Greiner & McGee, 2018;

Luke, 2019). Our hypotheses require that we consider potential inequalities in the *h*-index occurring within disciplines and also between them. As recommended given our substantive questions, first level variables were group mean centered (Enders & Tofighi, 2007). Group or cluster means centering is appropriate for questions focusing on within group differences.

All models were run using the lme4 package (Bates et al., 2007) in R version 1.4.1106.

Results

To better understand within and between discipline differences in the *h*-index, or Hirsch Index, we must first establish whether the multilevel approach adds to our understanding beyond a standard or fixed effects regression. One way to provide "empirical evidence of the need for multilevel modeling" is the intraclass correlation coefficient (ICC) (Luke, 2019, p. 18). Intraclass correlations are seen as the "first step" in multilevel modeling as they indicate the proportion of the total variance explained by the group level, indicating, here, disciplinary differences (see Lee, 2000). To observe the intraclass correlation, we begin with a null or fully unconditional model. The fully unconditional model of the *h*-index for the full dataset of high-performing scholars has an ICC of .23 meaning that 23% of the variability of the *h*-index occurs at the discipline level. This is consistent with the disciplinary differences hypothesis (hypothesis 3) that a significant portion of the variance in the *h*-index lies between disciplines.

Table 2 presents the multilevel models for the *h*-index. In models 1 and 2, we see the results for the complete data. The within-discipline effects are significant with the exception of the disciplinarity score. The percentage of female name is negative and significant, although the magnitude of the effect is modest. With each additional percentage above the disciplinary mean for the percentage female name, the *h*-index decreases by .01. This offers some evidence of the female penalty hypothesis (hypothesis 1). The largest effect is the percentage of sole author publications. With each additional percent increase beyond the disciplinary mean, the *h*-index declines by .28. Sole authorship has a significant effect on the *h*-index consistent with the sole author hypothesis (hypothesis 2). Other than disciplinarity, the control variables are positive and significant consistent with expectations one may have about cumulative advantage in academic work (Merton, 1968).

<Table 2 about here>

Figure 1 presents the estimated means for a distribution of 15 disciplines. We constructed the estimated mean by subtracting the conditional group means from the intercept for the full model (Table 2.2). This provides further evidence of the significant differences between disciplines. The Law is the discipline with the lowest estimated *h*-index with a score of 22.1, while Immunology and Epidemiology have the highest scores: 67.6 and 68.4, respectively.

<Figure 1 about here>

While the *h*-index is likely used by some evaluators without consideration of the many differences across fields, the full breadth of the high-performing scholar dataset runs the risk of apples-to-oranges comparison, alongside problems associated with the Scopus database. To evaluate whether similar processes occur within a more bounded field, we examine within and between discipline differences in the *h*-index for the field of clinical medicine. The unconditional model of the *h*-index for the clinical medicine subset has an ICC of .16. A smaller percentage of the variability is explained by disciplinary differences for this subset, but the level 2 effects remain moderately strong. Table 2, Models 3 and 4 present the multilevel models of the *h*-index for the clinical medicine subset. The effects are similar across the full models (comparing models 2 and 4) with the exception that the control variable for disciplinarity score is significant and positive in the clinical medicine model. The percentage female name has the same relationship with the *h*-index for the subset: negative and modest. The magnitude for the percentage of sole author publications is somewhat stronger, which may be expected given that sole authorship is rarer in clinical medicine. Each additional percentage beyond the within discipline mean decreases the *h*-index by .34.

<Figure 2 about here>

Like Figure 1, Figure 2 presents the estimated means for a distribution of 15 disciplines within the clinical medicine subset. Again, we see further evidence of the significant differences between disciplines. Dentistry and Otorhinolaryngology are the disciplines with the lowest scores - 39.8 and 39.9 - while epidemiology and immunology have the highest scores - 66.3 and 66.8.[3] These differences provide further evidence supporting the disciplinary differences hypothesis (hypothesis 3).

Conclusion

The *h*-index, or Hirsch Index, is a widely used metric used for performance evaluation or quality valuation in the sciences and across the academy. This research aimed to contribute to the literature on bibliometrics and inequality in science by examining both within and between discipline differences in the *h*-index. We used multilevel models predicting the *h*-index for high-performing scholars in 140 disciplines. Results indicate that gender and sole-authorship affect the *h*-index providing support for our female penalty and sole authorship hypotheses. The ICC score provided evidence of between discipline differences alongside pronounced differences in the distribution of the estimated mean supporting the disciplinary differences hypothesis. The robustness of these results were evaluated using a subset of clinical medicine scholars to reduce apples to oranges comparisons and the effects of idiosyncrasies in the Scopus database. These results were similar to the all disciplines models. In sum, important differences exist within and between disciplines in the *h*-index.

[3] Epidemiology is classified across multiple fields, including the social sciences and clinical medicine.

Data limitations suggest several avenues for future research. First, the data select on high-performing scholars, and, therefore do not generalize to the broader population of academics. While we believe that this likely results in a conservative estimate of within and between discipline inequality - a random sample of scholars is likely to show even more inequality on these dimensions - more data on academia writ large are required to verify this claim. Second, the estimation of gender suffers from well-known limitations. Data linking self-reported gender beyond the gender binary to citation data would be a welcome resource. Third, and along similar lines, most scholarship on publication and inequality focus on gender as it is possible, despite these weaknesses to infer gender from names and other aspects of inequality, like race and class, remain under-studied. Future research would benefit from data linking self-reported race and class to large bibliometric data to better understand the broad effects of inequality in academic work.

In conclusion, we believe that any debate about the merits of using metrics in performance evaluation in academia is resolved by this analysis and the significant body of research that proceeds it: Metrics provide an inadequate summary of academic careers and, moreover, are subject to within and between discipline biases. This analysis provided evidence of how the *h*-index disadvantages women scientists, but there is little reason to believe that the bias embedded in scholarly metrics do not extend to other dimensions of power, including race, class, and sexual orientation. These metrics simply should not be used for evaluating academic labor. As Cassidy Sugimoto succinctly states, "Ranking scientists is not good for science" (quoted in Van Noorden & Singh Chawla, 2019).

References

Akbaritabar, A., & Squazzoni, F. (2021). Gender Patterns of Publication in Top Sociological Journals. *Science, Technology, & Human Values*, *46*(3), 555–576. https://doi.org/10.1177/0162243920941588

Barnes, C. (2017). The h-index Debate: An Introduction for Librarians. *The Journal of Academic Librarianship*, *43*(6), 487–494. https://doi.org/10.1016/j.acalib.2017.08.013

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R Package Version*, *2*(1), 74.

Berg, L. D., Huijbens, E. H., & Larsen, H. G. (2016). Producing anxiety in the neoliberal university. *The Canadian Geographer / Le Géographe Canadien*, *60*(2), 168–180. https://doi.org/10.1111/cag.12261

Blevins, C., & Mullen, L. (2015). Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction. *DHQ: Digital Humanities Quarterly*, *9*(3).

Boston University. (2018). *Boston University School of Public Health Faculty Handbook*. http://www.bu.edu/sph/files/2019/01/BUSPH-Faculty-Handbook.-112918.pdf

Cech, E., Rubineau, B., Silbey, S., & Seron, C. (2011). Professional Role Confidence and Gendered Persistence in Engineering. *American Sociological Review*, *76*(5), 641–666. https://doi.org/10.1177/0003122411420815

Cole, J., & Zuckerman, H. (1984). The Productivity Puzzle. *Advances in Motivation and Achievement*, 217–258.

Edwards, M. A., & Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, *34*(1), 51–61. https://doi.org/10.1089/ees.2016.0223

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121.

Forrester, N. (2021). Mental health of graduate students sorely overlooked. *Nature*, *595*(7865), 135–137. https://doi.org/10.1038/d41586-021-01751-z

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Fox, M. F. (2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science*, *35*(1), 131–150.

Fox, M. F. (2020). Gender, science, and academic rank: Key issues and approaches. *Quantitative Science Studies*, *1*(3), 1001–1006. https://doi.org/10.1162/qss_a_00057

Fox, M. F., Whittington, K., & Linkova, M. (2017). Gender,(in) equity, and the scientific workforce. *Handbook of Science and Technology Studies*, 701–731.

Gardner, S. K. (2009). Conceptualizing Success in Doctoral Education: Perspectives of Faculty in Seven Disciplines. *The Review of Higher Education*, *32*(3), 383–406. https://doi.org/10.1353/rhe.0.0075

Grant, L., Kennelly, I., & Ward, K. B. (2000). Revisiting the gender, marriage, and parenthood puzzle in scientific careers. *Women's Studies Quarterly*, *28*(1/2), 62–85.

Greiner, P. T., & McGee, J. A. (2018). Divergent Pathways on the Road to Sustainability: A Multilevel Model of the Effects of Geopolitical Power on the Relationship between Economic Growth and Environmental Quality. *Socius*, *4*, 2378023117749381. https://doi.org/10.1177/2378023117749381

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hofstra, B., Kulkarni, V. V., Galvez, S. M.-N., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291.

Ioannidis, J. P. A., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLOS Biology*, *17*(8), e3000384. https://doi.org/10.1371/journal.pbio.3000384

Ioannidis, J. P. A., Boyack, K. W., & Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLOS Biology*, *18*(10), e3000918. https://doi.org/10.1371/journal.pbio.3000918

Kamrani, P., Dorsch, I., & Stock, W. G. (2021). Do researchers know what the h-index is? And how do they estimate its importance? *Scientometrics*, *126*(7), 5489–5508. https://doi.org/10.1007/s11192-021-03968-1

Katz, Y., & Matter, U. (2020). Metrics of Inequality: The Concentration of Resources in the U.S. Biomedical Elite. *Science as Culture*, *29*(4), 475–502. https://doi.org/10.1080/09505431.2019.1694882

Kelly, C. D., & Jennions, M. D. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution*, *21*(4), 167–170. https://doi.org/10.1016/j.tree.2006.01.005

Siler, K., & Larivière, V. (2022). Who games metrics and rankings? Institutional niches and journal impact factor inflation. *Research Policy*, *51*(10), 104608.

King, M. M., & Frederickson, M. E. (2021). The Pandemic Penalty: The Gendered Effects of COVID-19 on Scientific Productivity. *Socius*, *7*, 23780231211006976. https://doi.org/10.1177/23780231211006977

Lanahan, L., Graddy-Reed, A., & Feldman, M. P. (2016). The Domino Effects of Federal Research Funding. *PLOS ONE*, *11*(6), e0157325. https://doi.org/10.1371/journal.pone.0157325

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender

disparities in science. *Nature*, *504*(7479), 211–213.
Leahey, E. (2006). Gender Differences in Productivity: Research Specialization as a Missing Link. *Gender & Society*, *20*(6), 754–780. https://doi.org/10.1177/0891243206293030
Leahey, E. (2007). Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings. *American Sociological Review*, *72*(4), 533–561. https://doi.org/10.1177/000312240707200403
Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, *35*(2), 125–141.
Light, R. (2009). Gender Stratification and Publication in American Science: Turning the Tools of Science Inward. *Sociology Compass*, *3*, 721–733. https://doi.org/10.1111/j.1751-9020.2009.00221.x
Light, R. (2013). Gender Inequality and the Structure of Occupational Identity: The Case of Elite Sociological Publication. *Research in the Sociology of Work*, *24*, 239–268.
Long, J. S., & Fox, M. F. (1995). Scientific careers: Universalism and particularism. *Annual Review of Sociology*, *21*(1), 45–71.
Luke, D. A. (2019). *Multilevel modeling*. Sage publications.
Ma, Y., & Xiao, S. (2021). Math and Science Identity Change and Paths into and out of STEM: Gender and Racial Disparities. *Socius*, *7*, 23780231211001976. https://doi.org/10.1177/23780231211001978
McKiernan, E. C., Schimanski, L. A., Nieves, C. M., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Meta-research: Use of the journal impact factor in academic review, promotion, and tenure evaluations. *Elife*, *8*, e47338.
Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.
Mihaljević, H., Tullney, M., Santamaría, L., & Steinfeldt, C. (2019). Reflections on gender analyses of bibliographic corpora. *Frontiers in Big Data*, *2*, 29.
Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5
Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.
Morgan, A. C., Way, S. F., Hoefer, M. J. D., Larremore, D. B., Galesic, M., & Clauset, A. (2021). The unequal impact of parenthood in academia. *Science Advances*, *7*(9), eabd1996. https://doi.org/10.1126/sciadv.abd1996
Muller, J. Z. (2019). *The tyranny of metrics*. Princeton University Press.
Niles, M. T., Schimanski, L. A., McKiernan, E. C., & Alperin, J. P. (2020). Why we publish where we do: Faculty publishing values and their relationship to review, promotion and tenure expectations. *PLOS ONE*, *15*(3), e0228914. https://doi.org/10.1371/journal.pone.0228914
Ohio State University. (2014). *Department of Surgergy, Appointments, Promotion, and Tenure*. https://oaa.osu.edu/sites/default/files/uploads/governance-documents/college-of-medicine/surgery/Surgery_APT_5-20-14.pdf
Oregon State University. (2020). *Oregon State University College of Business Faculty Handbook*. https://business.oregonstate.edu/sites/default/files/cob-faculty-handbook-06-2020.pdf
Parish, A. J., Boyack, K. W., & Ioannidis, J. P. A. (2018). Dynamics of co-authorship and productivity across different fields of scientific research. *PLOS ONE*, *13*(1), e0189742. https://doi.org/10.1371/journal.pone.0189742
Penner, A. M., & Willer, R. (2019). Men's Overpersistence and the Gender Gap in Science and Mathematics. *Socius*, *5*, 2378023118821836. https://doi.org/10.1177/2378023118821836
Purvis, A. (2006). The h index: Playing the numbers game. *Trends in Ecology & Evolution*, *21*(8), 422. https://doi.org/10.1016/j.tree.2006.05.014
Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of

Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142.
Stephan, P. (2012). *How economics shapes science*. Harvard University Press.
Teele, D. L., & Thelen, K. (2017). Gender in the journals: Publication patterns in political science. *PS: Political Science & Politics*, *50*(2), 433–447.
Van Noorden, R., & Singh Chawla, D. (2019). Hundreds of extreme self-citing scientists revealed in new database. *Nature*, *572*(7771), 578–579. https://doi.org/10.1038/d41586-019-02479-7
Xie, Y., & Shauman, K. A. (1998). Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review*, *63*(6), 847–870. https://doi.org/10.2307/2657505

Table 1: Descriptive Statistics

	All Disciplines			Clinical Medicine Subset		
Statistic	N	Mean	St. Dev.	N	Mean	St. Dev.
h-index	42,509	45.6	21	16,041	53.2	22.1
% Female Name	42,509	19.1	37.7	16,041	19.7	38.4
% Sole Author Publications	42,509	9.8	12.3	16,041	8.2	9.1
# of Citations	42,509	10,876.80	12,910.80	16,041	14,005.30	14,971.50
University Count	42,509	198.3	203.1	16,041	182	200.2
% Disciplinarity	42,509	73.1	19.2	16,041	79.9	16.9
Age	42,509	35.5	9.9	16,041	36.1	9.2

Table 2: A Multilevel Model of the *h*-index

	All Disciplines		Clinical Medicine Subset	
	Null Model	Full Model	Null Model	Full Model
	(1)	(2)	(3)	(4)
% Female Name		-0.01***		-0.01***
% Sole Author Publications		-0.28***		-0.34***
# of Citations (1996-2019)		0.12***		0.12***
University Count (# of Citation Allstars)		0.004***		0.003***
Disciplinarity Score		-0.001		0.02***
Age (Years Publishing)		0.18***		0.17***
Constant	41.08***	40.94***	51.00***	49.30***
ICC	.23		.16	
N	42,509	42,509	16,041	16,041
Log Likelihood	-183,302.70	-154,035.40	-71,248.87	-58,339.33
AIC	366,611.30	308,088.70	142,503.70	116,696.70
BIC	366,637.30	308,166.70	142,526.80	116,765.80

*p < .05; **p < .01; ***p < .001

Data: Ioannidis et al. (2019)

Figure 1: Distribution of Estimated Disciplinary *h*-index Means (All Disciplines)

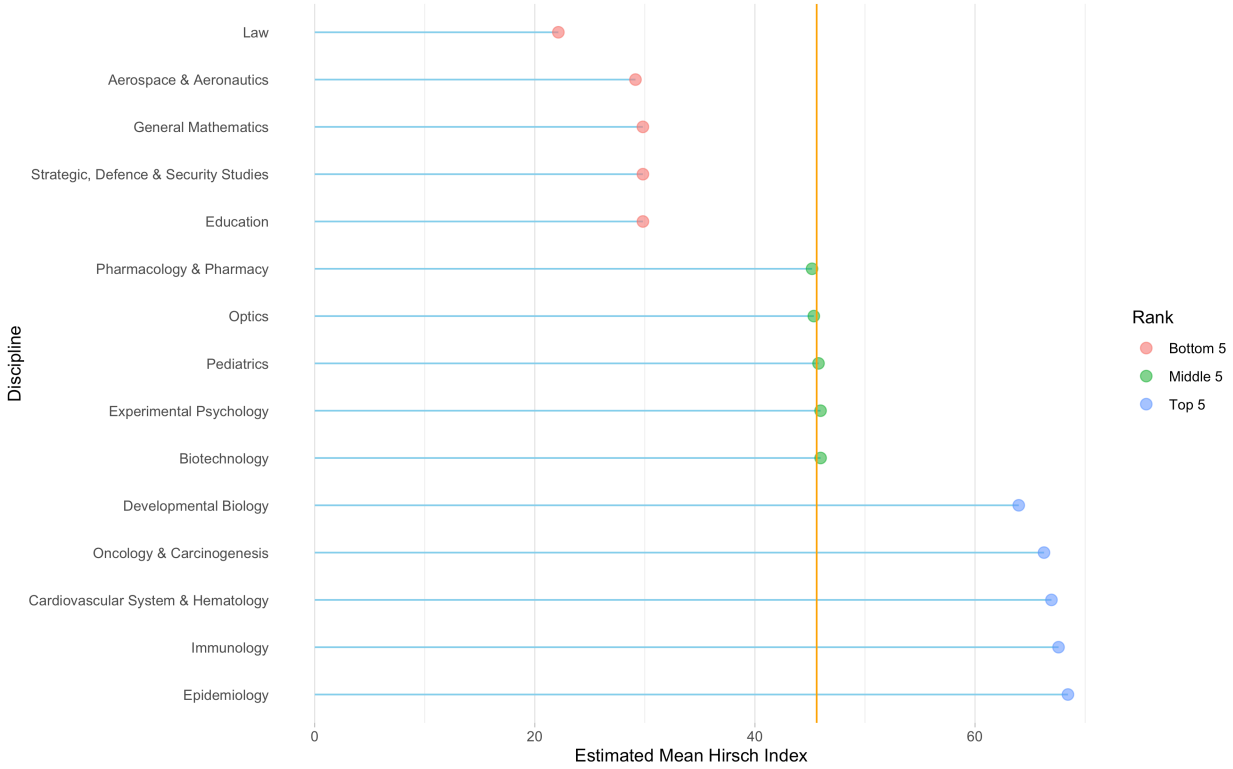

Figure 2: Distribution of Estimated Disciplinary *h*-index Means (Clinical Medicine Subset)