

A fairer way to compare researchers at any career stage and in any discipline using open-access citation data

Corey J. A. Bradshaw^{1,2}, Justin M. Chalker³, Stefani A. Crabtree^{4,5,6}, Bart A. Eijkelkamp⁷, John A. Long⁷, Justine R. Smith⁸, Kate Trinajstić⁹, and Vera Weisbecker^{2,7}

¹Global Ecology, College of Science and Engineering, Flinders University

²ARC Centre of Excellence for Australian Biodiversity and Heritage

³Institute for Nanoscale Science and Technology, College of Science and Engineering, Flinders University

⁴Department of Environment and Society, Utah State University

⁵The Santa Fe Institute

⁶The Centre for Research and Interdisciplinarity

⁷College of Science and Engineering, Flinders University

⁸Eye and Vision Health, Flinders University College of Medicine and Public Health, Flinders University

⁹School of Molecular and Life Sciences, Curtin University

October 26, 2020

Abstract

The pursuit of simple, yet fair, unbiased, and objective measures of researcher performance has occupied bibliometricians and the research community as a whole for decades. However, despite the diversity of available metrics, most are either complex to calculate or not readily applied in the most common assessment exercises (e.g., grant assessment, job applications). The ubiquity of metrics like the h -index (h papers with at least h citations) and its time-corrected variant, the m -quotient (h -index \div number of years publishing) therefore reflect the ease of use rather than their capacity to differentiate researchers fairly among disciplines, career stage, or gender. We address this problem here by defining an easily calculated index based on publicly available citation data (Google Scholar) that corrects for most biases and allows assessors to compare researchers at any stage of their career and from any discipline on the same scale. Our ϵ' -index violates fewer statistical assumptions relative to other metrics when comparing groups of researchers, and can be easily modified to remove inherent gender biases in citation data. We demonstrate the utility of the ϵ' -index using a sample of 480 researchers with Google Scholar profiles, stratified evenly into eight disciplines (archaeology, chemistry, ecology, evolution and development, geology, microbiology, ophthalmology, palaeontology), three career stages (early, mid-, late-career), and two genders. We advocate the use of the ϵ' -index whenever assessors must compare research performance among researchers of different backgrounds, but emphasise that no single index should be used exclusively to rank researcher capability.

Introduction

Deriving a fair, unbiased, and easily generated quantitative index serving as a reasonable first-pass metric for comparing the relative performance of academic researchers is — by the very complexity, diversity, and intangibility of research output across academic disciplines — impossible (1). However, that unachievable

aim has not discouraged bibliometricians and non-bibliometricians alike from developing scores of citation-based variants (2, 3) in an attempt to do exactly that, from the better-known h -index (4, 5) (h papers with at least h citations), m -quotient (4, 5) (h -index \div number of years publishing), and g -index (6) (unique largest number such that the top g papers decreasingly ordered by citations have least g^2 citations), to the scores of variants of these and other indices — e.g., h^2 -index, e -index (7), χ -index (8), h_m -index (9), g_m -index (10), etc. (3). Each metric has its own biases and strengths (11-13), suggesting that several should be used simultaneously to assess citation performance. For example, the arguably most-popular h -index down-weights quality relative to quantity (14), ignores the majority of accumulated citations in the most highly cited papers (15), has markedly different distributions among disciplines (16), and tends to increase with experience (17). The h -index can even rise following the death of the researcher, because the h -index can never decline (2) and citations can continue to accumulate posthumously.

Despite their broad use in everything from assessing candidates applying for academic positions, comparing the track records of researchers applying for grants, to applications for promotion (3, 18) *inter alia*, single-value citation metrics are rarely meant to (nor should they) be definitive assessment tools (3). Instead, their most valuable (and fair) application is to provide a quick ‘first pass’ to rank a sample of researchers, followed by more detailed assessment of publication quality, experience, grant successes, mentorship, collegiality and all the other characteristics that make a researcher more or less competitive for rare positions and grant monies. But despite the many different metrics available and arguable improvements that have been proposed since 2005 when the h -index was first developed (4, 5), few are used regularly in these regards. This is because they are difficult to calculate without detailed data of a candidate’s publication history, they are not readily available on open-access websites, and/or they tend to be highly correlated with the h -index anyway (19). It is for these reasons that the admittedly flawed (20, 21) h -index and its experienced-corrected variant, the m -quotient, are still the dominant (h -index much more so than the m -quotient) (2) metrics employed given that they are easily calculated (2, 22) and found for most researchers on open-access websites such as Google Scholar (23) (scholar.google.com). The lack of access and detailed understanding of the many other citation-based metrics means that most of them go unused (3), and are essentially valueless for everyday applications of researcher assessment.

The specific weaknesses of the h -index or m -quotient make the comparison of researchers in different career stages, genders, and disciplines unfair because they are not normalised in any way. Furthermore, there is no quantitatively supported threshold above or below which assessors can easily ascertain minimum citation performance for particular applications — while assessors certainly use subjective ‘rules of thumb’, a more objective approach is preferable. For this reason, an ideal citation-based metric should only be considered as a relative index of performance, but relative to what, and to whom?

To address these issues and to provide assessors with an easy, rapid, yet objective *relative* index of citation performance for any group of researchers, we designed a new index we call the ‘ ϵ -index’ (the ‘ ϵ ’ signifies the use of residuals, or deviance from a trend) that is simple to construct, can be standardised across disciplines, is meaningful only as a relative index for a particular sample of researchers, can be corrected for career breaks (see Methods), and provides a sample-specific threshold above and below which assessors can determine whether individual performance is greater or less than that expected relative to the other researchers in the specific sample.

With the R code we provide, an assessor need only acquire four separate items of information from Google Scholar (or if they have access, from other databases such as Scopus — scopus.com) to calculate a researcher’s ϵ -index: (i) the number of citations acquired for the researcher’s top-cited paper (i.e., the first entry in the Google Scholar profile), (ii) the $i10$ -index (number of articles with at least 10 citations), (iii) the h -index, and (iv) the year in which the researcher’s first peer-reviewed paper was published. While the last item requires sorting a researcher’s outputs by year and scrolling to the earliest paper, this is not a time-consuming process. We demonstrate the performance of the ϵ -index using Google Scholar citation data we collected for 480 researchers in eight separate disciplines spread equally across genders and career stages to show how the index performs relative to the m -quotient (the only other readily available, opportunity-

corrected citation index available on Google Scholar) across disciplines, career stages, and genders. We also provide a simple method to scale the index across disciplines (ϵ' -index) to make researchers in different areas comparable despite variable citation trends within their respective areas.

Results

Despite the considerable variation in citation metrics among researchers and disciplines, there was broad consistency in the strength of the relationships between citation mass (A_{rel}) and \log_e years publishing (t) across disciplines (Fig. 1), although the geology (GEO) sample had the poorest fit ($^{\text{ALL}}R^2 = 0.43$; Fig. 1). The distribution of residuals ϵ for each discipline revealed substantial difference in general form and central tendency (Fig. 2), but after scaling, the distributions of ϵ' became aligned among disciplines and were approximately Gaussian (Shapiro-Wilk normality tests; see Fig. 2 for test values).

After scaling (Fig. 3a), the relationship between ϵ' and the m -quotient is non-linear and highly variable (Fig. 3b), meaning that m -quotients often poorly reflect actual relative performance (and despite the m -quotient already being ‘corrected’ for t , it still increases with t ; Supplementary Material Fig. S1). For example, there are many researchers whose m -quotient < 1 , but who perform above expectation ($\epsilon' > 0$). Alternatively, there are many researchers with an m -quotient of up to 2 or even 3 who perform below expectation ($\epsilon' < 0$). Once the m -quotient > 3 , ϵ' reflects above-expectation performance for all researchers in the example sample (Fig. 3b). The corresponding ϵ' indicate a more uniform spread by gender and career stage (Fig. 3c) than do m -quotients (Fig. 3d). Another advantage of ϵ' *versus* the m -quotient is that the former has a threshold ($\epsilon' = 0$) above which researchers perform above expectation and below which they perform below expectation, whereas the m -quotient has no equivalent threshold. Further, the m -quotient tends to increase through one’s career, whereas ϵ' is more stable. There is still an increase in ϵ' during late career relative to mid-career, but this is less pronounced than that observed for the m -quotient (Fig. 4).

Examining the ranks derived from ϵ' across disciplines, genders and career stage (Fig. 5), bootstrapped median ranks overlap for all disciplines (Fig. 5a), but there are some notable divergences between the genders across career stage (Fig. 5b). In general, women ranked slightly below men in all career stages, although the bootstrapped median ranks overlap among early and mid-career researchers. However, the median ranks for late-career women and men do not overlap (Fig. 5b), which possibly reflects the observation that senior academic positions in many disciplines are dominated by men (24-26), and that women tend to receive fewer citations than men at least in some disciplines, which often tends to compound over time (27-30). The ranking based on the m -quotient demonstrates the disparity among disciplines (Fig. 5c), but it is perhaps somewhat more equal between the genders (Fig. 5d) compared to the ϵ' rank (Fig. 5b), despite the higher variability of the m -quotient bootstrapped median rank.

However, calculating the scaled residuals across all disciplines for each gender separately, and then combining the two datasets and recalculating the rank (producing a gender-‘debiased’ rank) effectively removed the gender differences (Fig. 6).

Discussion

Todeschini and Baccini (31) recommended that the ideal author-level indicator of citation performance should (i) have an unequivocal mathematical definition, (ii) be easily computed from available data (for a detailed breakdown of implementation steps and the R code function, see github.com/cjabradshaw/EpsilonIndex), (iii) balance rankings between more experienced and novice researchers (iv) while preserving sensitivity to the performance of top researchers, and (iv) be sensitive to the number and distribution of citations and articles. Our new ϵ -index not only meets these criteria, it also adds the ability to compare across disciplines by using a simple scaling approach, and can easily be adjusted for career gaps by subtracting research-inactive periods from the total number of years publishing (t). In this way, the ϵ -index could prove invaluable as we move toward greater interdisciplinarity, where tenure committees have had difficulty assessing the performance of candidates straddling disciplines (32, 33). The ϵ -index does not ignore high-citation papers, but neither

does it overemphasise them, and it includes an element of publication frequency (i_{10}) while simultaneously incorporating an element of ‘quality’ by including the h -index.

Like all other existing metrics, the ϵ -index does have some disadvantages in terms of not correcting for author contribution — such as the h_m -index (9) or g_m -index (10) — even though these types of metrics can be cumbersome to calculate. Early career researchers who have published but have yet to be cited will not yet be able to calculate their ϵ -index, as they will not have an h -index score, so would require different types of assessment. Another potential limitation is that the ϵ -index alone does not correct for any systemic gender biases associated with the many reasons why women tend to be cited less than men (24-30, 34), but it does easily allow an assessor to benchmark any subset of researchers (e.g., women-only or men-only) to adjust the threshold accordingly. Thus, women can be compared to other women and ranked accordingly such that the ranks are more comparable between these two genders. Alternatively, dividing the genders and benchmarking them separately followed by a combined re-ranking (Fig. 6) effectively removes the gender bias in the ϵ -index, which is difficult or impossible to do with other ranking metrics. We certainly advocate this approach when assessing mixed-gender samples (the same approach could be applied to other subsets of researchers deemed *a priori* to be at a disadvantage).

The ϵ -index also potentially suffers from the requirement of the constituent citation data upon which it is based being accurate and up-to-date (35, 36). Regardless, should an assessor have access to potentially more rigorous citation databases (e.g., Scopus), the ϵ -index can still be readily calculated, although within-sample consistency must be maintained for the ranks to be meaningful. We also show that the distribution of the ϵ -index is relatively more Gaussian in behaviour than the time-corrected m -quotient, with the added advantage of identifying a threshold above and below which individuals are deemed to be performing better or worse than expected relative to their sample peers. While there are potentially subjective rules of thumb for thresholds to be applied to the m -quotient, the residual nature of the ϵ -index makes it a more objective metric for assessing relative rank, and the ϵ -index is less-sensitive than the m -quotient regarding the innate rise of ranking as a researcher progresses through her career (Fig. 4).

We reiterate that while the ϵ -index is an advance on existing approaches to rank researchers according to their citation history, a single metric should never be the sole measure of a researcher’s productivity or potential. Nonetheless, the objectivity, ease of calculation, and flexibility of its application argue that the ϵ -index is a needed tool in the quest to provide fair initial appraisals of a researcher’s publication performance.

Materials and Methods

Researcher samples

Each co-author assembled an example set of researchers from within her/his field, which we broadly defined as *archaeology* (S.A.C.), *chemistry* (J.M.C.), *ecology* (C.J.A.B.), *evolution/development* (V.W.), *geology* (K.T.), *microbiology* (B.A.E.), *ophthalmology* (J.R.S.), and *palaeontology* (J.A.L.). Our basic assembly rules for each of these discipline samples were: (i) 20 researchers from each stage of career, defined here arbitrarily as *early career* (0–10 years since first peer-reviewed article published in a recognised scientific journal), *mid-career* (11–20 years since first publication), and *late career* (> 20 years since first publication); each discipline therefore had a total of 60 researchers, for a total sample of $8 \times 60 = 480$ researchers across all disciplines. (ii) Each sample had to include an equal number of women and men from each career stage. (iii) Each researcher had to have a unique, publicly accessible Google Scholar profile with no obvious errors, inappropriate additions, obvious omissions, or duplications. The entire approach we present here assumes that each researcher’s Google Scholar profile is accurate, up-to-date, and complete.

We did not impose any other rules for sample assembly, but encouraged each compiler to include only a few previous co-authors. Our goal was to have as much ‘inside knowledge’ as possible with respect to each discipline, but also to include a wide array of researchers who were predominantly independent of each of us. The composition of each sample is somewhat irrelevant for the purposes of our example dataset; we merely attempted gender and career-level balance to show the properties of the ranking system (i.e., we did not

intend for sampling to be a definitive comment about the performance of particular researchers, nor did we mean for each sample to represent an entire discipline). Finally, we completely anonymised the sample data for publication.

Citation data

Our overall aim was to provide a meaningful and objective method for ranking researchers by citation history without requiring extensive online researching or information that was not easily obtainable from a publicly available, online profile. We also wanted to avoid an index that was overly influenced by outlier citations, while still keeping valuable performance information regarding high-citation outputs and total productivity (number of outputs).

For each researcher, the algorithm requires the following information collected from Google Scholar: (i) **i_{10} -index** (the number of publications in the researcher’s profile with at least 10 citations, which we denoted i_{10}); one condition is that a researcher must have $i_{10} \geq 1$ for the algorithm to function correctly; (ii) **h -index** — the researcher’s Hirsch number (4): the number of publications with at least as many citations, which we denoted h ; (iii) the number of citations for the researcher’s most highly cited paper (denoted c_m); and (iv) the year the researcher published her/his first peer-reviewed article in a recognized scientific journal (denoted Y_1). For the designation of Y_1 , we excluded any reports, chapters, books, theses or other forms of publication that preceded the year of the first peer-reviewed article; however, we included citations from the former sources in the researcher’s i_{10} , h , and c_m .

Ranking algorithm

The algorithm first computes a power-law-like relationship between the vector of frequencies (as measured from Google Scholar): i_{10} , h , and 1, and the vector of their corresponding values: 10, h , and c_m , respectively. Thus, h is, by definition, both a frequency (y -axis) and value (x -axis). We then calculated a simple linear model of the form $y \sim a + \beta x$, where

$$y = \log_e$$

$$\begin{bmatrix} i_{10} \\ h \\ 1 \end{bmatrix} \text{ and } x = \log_e$$

$$\begin{bmatrix} 10 \\ h \\ c_m \end{bmatrix}$$

(y is the citation frequency, and x is the citation value) for each researcher (Supplementary Material Fig. S2). The corresponding $\hat{\alpha}$ and $\hat{\beta}$ for each relationship allowed us to calculate a standardized integral (area under the power-law relationship, A_{rel}) relative to the researcher in the sample with the highest c_m . This implies all areas were scaled to the maximum in the sample.

A researcher’s A_{rel} therefore represents her/his citation *mass*, but this value still requires correction for individual opportunity (time since first publication, $t = \text{current year} - Y_1$) to compare researchers at different stages of their career. This is where career gaps can be taken into account explicitly for any researcher in the sample by subtracting a_i = the total cumulative time absent from research (e.g., maternity or paternity leave, sick leave, secondment, etc.) for individual i from t , such that an individual’s career gap-corrected $t' = t - a$. We therefore constructed another linear model of the form $A_{\text{rel}} \sim \gamma + \theta \log_e t$ across all researchers in the sample, and took the residual (ϵ) of an individual researcher’s A_{rel} from the predicted relationship as a metric of citation performance relative to the rest of the researchers in that sample (Supplementary Material Fig. S3). This residual ϵ allows us to rank all individuals in the sample from highest (highest citation performance relative to opportunity and the entire sample) to lowest (lowest citation performance relative to opportunity and the entire sample). Any researcher in the sample with a

positive ϵ is considered to be performing above expectation (relative to the group and the time since first publication), and those with a negative ϵ fall below expectation. This approach also has the advantage of fitting different linear models to subcategories within a sample to rank researchers within their respective groupings (e.g., such as by gender; Supplementary Material Fig. S4). An R code function to produce the index and its variants using a sample dataset is available from github.com/cjabradshaw/EpsilonIndex.

Discipline standardization

Each discipline has its own citation characteristics and trends (16), so we expect that the distribution of residuals (ϵ) within each discipline to be meaningful only for that discipline’s sample. We therefore endeavoured to scale (‘normalise’) the results such that researchers in different disciplines could be compared objectively and fairly.

We first scaled the A_{rel} within each discipline by dividing each i researcher’s A_{rel} by the sample’s root mean square:

$$A'_{\text{rel}_i} = \frac{A_{\text{rel}_i}}{\sqrt{\frac{\sum_{i=1}^n A_{\text{rel}_i}}{n-1}}}$$

where n = the total number of researchers in the sample ($n = 60$). We then regressed these discipline-scaled A'_{rel} against the \log_e number of years since first publication pooling all disciplines together, and then ranked these scaled residuals (ϵ') as described above.

Acknowledgments

We acknowledge our many peers for their stewardship of their online citation records. We acknowledge the Indigenous Traditional Owners of the land on which Flinders University is built — the Kaurna people of the Adelaide Plains.

Author Contributions

C.J.A.B conceptualised the paper, and C.J.A.B. designed the research. All co-authors provided sample data from their disciplines. C.J.A.B. wrote the code and did the analysis. C.J.A.B. wrote the paper, with contributions from all other authors.

ORCIDiDs

C.J.A. Bradshaw: 0000-0002-5328-7741; J.M. Chalker: 0000-0002-7504-5508; S.A. Crabtree: 0000-0001-8585-8943; B.A. Eijkelkamp: 0000-0003-0179-8977; J.A. Long: 0000-0001-8012-0114; J.R. Smith: 0000-0002-4756-5493; K. Trinajstić: 0000-0002-6519-6396; V. Weisbecker: 0000-0003-2370-4046

References

1. T. J. Phelan, A compendium of issues for citation analysis. *Scientometrics* **45**, 117-136 (1999).
2. C. Barnes, The h -index debate: an introduction for librarians. *J. Acad. Libr.* **43**, 487-494 (2017).
3. L. Wildgaard, “An overview of author-level indicators of research performance” in Springer Handbook of Science and Technology Indicators, W. Glänzel, H. F. Moed, U. Schmoch, M. Thelwall, Eds. (Springer International Publishing, Cham, 2019), 10.1007/978-3-030-02511-3_14, pp. 361-396.
4. J. E. Hirsch, An index to quantify an individual’s scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16569-16572 (2005).

5. A. Schubert, G. Schubert, *Äll along the h -index-related literature: a guided tour* in Springer Handbook of Science and Technology Indicators, W. Glänzel, H. F. Moed, U. Schmoch, M. Thelwall, Eds. (Springer International Publishing, Cham, 2019), 10.1007/978-3-030-02511-3_12, pp. 301-334.
6. L. Egghe, How to improve the h -index. *The Scientist* **20**, 15 (2006).
7. C.-T. Zhang, The e -Index, complementing the h -Index for excess citations. *PLoS One* **4**, e5429 (2009).
8. T. Fenner, M. Harris, M. Levene, J. Bar-Ilan, A novel bibliometric index with a simple geometric interpretation. *PLoS One* **13**, e0200098 (2018).
9. M. Schreiber, A modification of the h -index: The h_m -index accounts for multi-authored manuscripts. *J. Informetr.* **2**, 211-216 (2008).
10. M. Schreiber, How to modify the g -index for multi-authored manuscripts. *J. Informetr.* **4**, 42-54 (2010).
11. D. F. Thompson, E. C. Callen, M. C. Nahata, New indices in scholarship assessment. *Am. J. Pharm. Educ.* **73**, 111-111 (2009).
12. L. Bornmann, R. Mutz, H.-D. Daniel, Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *J. Am. Soc. Inf. Sci. Tec.* **59**, 830-837 (2008).
13. L. Bornmann, R. Mutz, S. E. Hug, H.-D. Daniel, A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *J. Informetr.* **5**, 346-359 (2011).
14. R. Costas, M. Bordons, The h -index: advantages, limitations and its relation with other bibliometric indicators at the micro level. *J. Informetr.* **1**, 193-203 (2007).
15. T. R. Anderson, R. K. S. Hankin, P. D. Killworth, Beyond the Durfee square: enhancing the h -index to score total publication output. *Scientometrics* **76**, 577-588 (2008).
16. P. D. Batista, M. G. Campiteli, O. Kinouchi, Is it possible to compare researchers with different scientific interests? *Scientometrics* **68**, 179-189 (2006).
17. C. D. Kelly, M. D. Jennions, The h index and career assessment by numbers. *Trends Ecol. Evol.* **21**, 167-170 (2006).
18. J. E. Hirsch, Does the h index have predictive power? *Proc. Natl. Acad. Sci. USA* **104**, 19193 (2007).
19. L. Bornmann, Redundancies in h index variants and the proposal of the number of top-cited papers as an attractive indicator. *Measurement* **10**, 149-153 (2012).
20. R. Costas, T. Franssen, Reflections around ‘the cautionary use’ of the h -index: response to Teixeira da Silva and Dobránszki. *Scientometrics* **115**, 1125-1130 (2018).
21. G. Abramo, C. A. D’Angelo, F. Viel, The suitability of h and g indexes for measuring the research performance of institutions. *Scientometrics* **97**, 555-570 (2013).
22. Y. Bhattacharjee, Impact factor. *Science* **309**, 1181 (2005).
23. E. Delgado López-Cózar, E. Orduña-Malea, A. Martín-Martín, “Google Scholar as a data source for research assessment” in Springer Handbook of Science and Technology Indicators, W. Glänzel, H. F. Moed, U. Schmoch, M. Thelwall, Eds. (Springer International Publishing, Cham, 2019), 10.1007/978-3-030-02511-3_4, pp. 95-127.
24. T. Tregenza, Gender bias in the refereeing process? *Trends Ecol. Evol.* **17**, 349-350 (2002).
25. V. Larivière, C. Ni, Y. Gingras, B. Cronin, C. R. Sugimoto, Global gender disparities in science. *Nature* **504**, 211-213 (2013).
26. L. Howe-Walsh, S. Turnbull, Barriers to women leaders in academia: tales from science and technology. *Stud. High. Educ.* **41**, 415-428 (2016).
27. D. W. Aksnes, Characteristics of highly cited papers. *Res. Eval.* **12**, 159-170 (2003).
28. D. Maliniak, R. Powers, B. F. Walter, The gender citation gap in international relations. *Intl. Organ.* **67**, 889-922 (2013).
29. C. Beaudry, V. Larivière, Which gender gap? Factors affecting researchers’ scientific impact in science and medicine. *Res. Policy* **45**, 1790-1817 (2016).
30. A. L. Atchison, Negating the gender citation advantage in political science. *PS-Polit. Sci. Polit.* **50**, 448-455 (2017).
31. R. Todeschini, A. Baccini, *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research* (Wiley-VCH, Weinheim, 2016).

32. J. Austin, Interdisciplinarity and tenure. *Science* **10 January** (2003).
33. E. Evans, *Paradigms, Interdisciplinarity, and Tenure*. PhD thesis (Stanford University, Palo Alto, California, USA, 2016), ISBN: 978-0-4386-6106-6, pp. 83.
34. T. E. Carter, T. E. Smith, P. J. Osteen, Gender comparisons of social work faculty using H-Index scores. *Scientometrics* **111**, 1547-1557 (2017).
35. J. A. Teixeira da Silva, J. Dobránszki, Multiple versions of the h -index: cautionary use for formal academic purposes. *Scientometrics* **115**, 1107-1113 (2018).
36. J. A. Teixeira da Silva, J. Dobránszki, Rejoinder to “Multiple versions of the h -index: cautionary use for formal academic purposes”. *Scientometrics* **115**, 1131-1137 (2018).

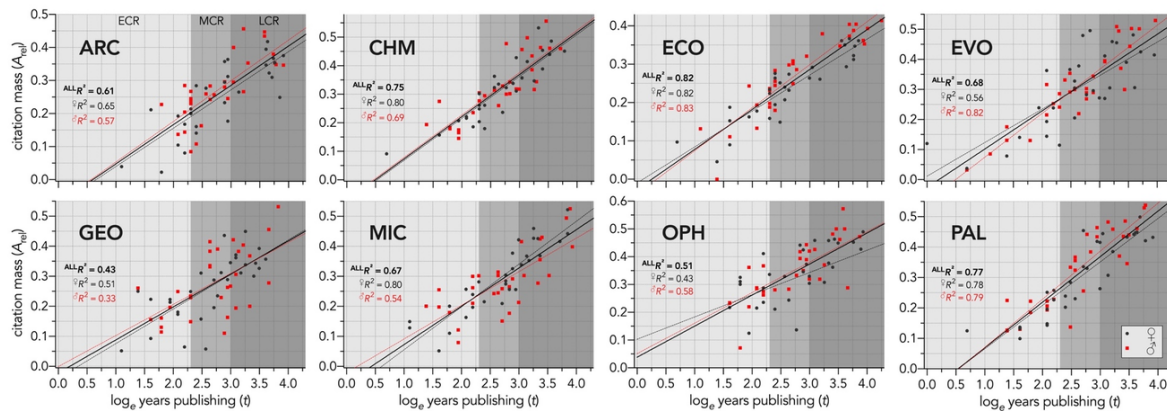


Figure 1: **Figure 1** | Relationship between a researcher's citation mass (A_{rel} ; area under the citation frequency-value curve — see Supplementary Material Fig. S2) and \log_e years (t) since first peer-reviewed publication (Y_1) for eight disciplines (ARC = archaeology, CHM = chemistry, ECO = ecology, EVO = evolution and development, GEO = geology, MIC = microbiology, OPH = ophthalmology, PAL = paleontology) comprising 60 researchers each (30 female, 30 male) in three different career stages: early career researcher (ECR), mid-career researcher (MCR), and late career researcher (LCR). The fitted lines correspond to the entire sample (solid black), women only (dashed black), and men only (dashed red). Information-theoretic evidence ratios for all relationships > 180 ; adjusted R^2 for each relationship shown in each panel.

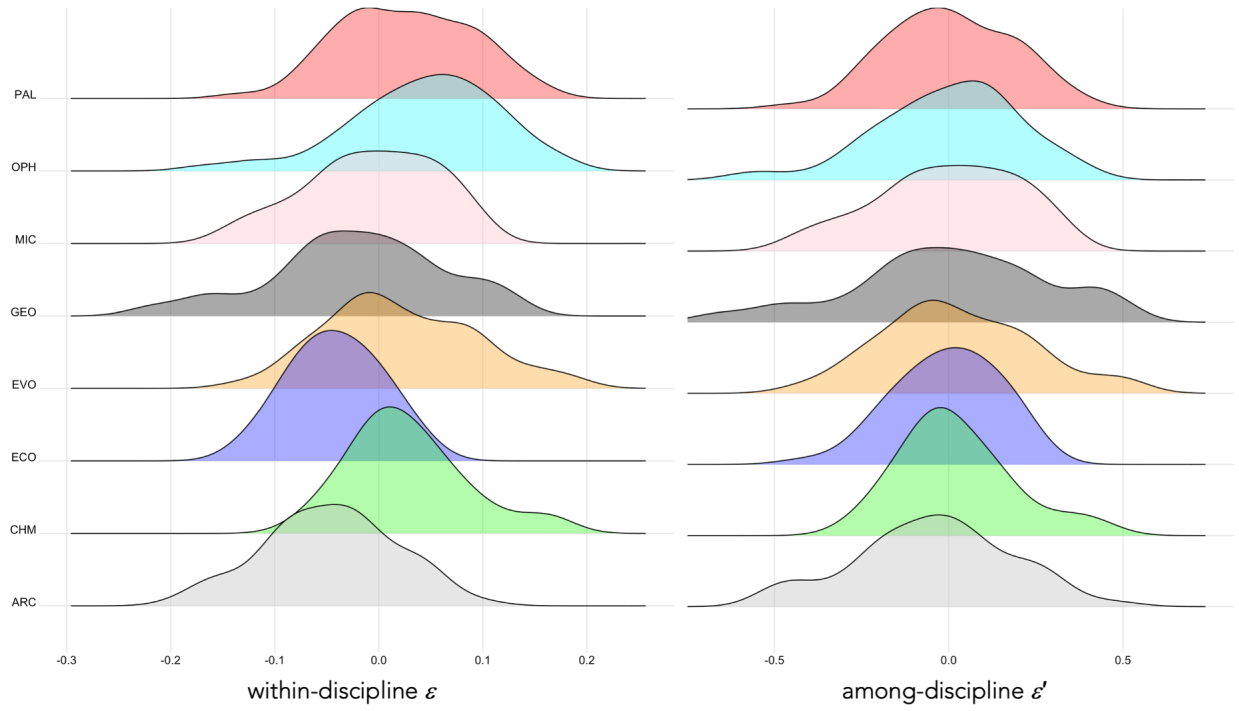


Figure 2: **Figure 2** | Left panel: Distribution of within-discipline residuals (ϵ) of the relationship between A_{rel} and \log_e years publishing (t) by discipline (ARC = archaeology, CHM = chemistry, ECO = ecology, EVO = evolution and development, GEO = geology, MIC = microbiology, OPH = ophthalmology, PAL = palaeontology), each comprising 60 researchers (30 female, 30 male). Right panel: Distribution of among-discipline residuals (ϵ') of the relationship between A'_{rel} (scaled) and t by discipline. All A'_{rel} distributions are approximately Gaussian behaviour according to Shapiro-Wilk normality tests (ARC: $W = 0.985$, $p = 0.684$; CHM: $W = 0.961$, $p = 0.051$; ECO: $W = 0.980$, $p = 0.409$; EVO: $W = 0.984$, $p = 0.630$; GEO: $W = 0.929$, $p = 0.398$; MIC: $W = 0.971$, $p = 0.170$; OPH: $W = 0.980$, $p = 0.416$; PAL: $W = 0.986$, $p = 0.720$).

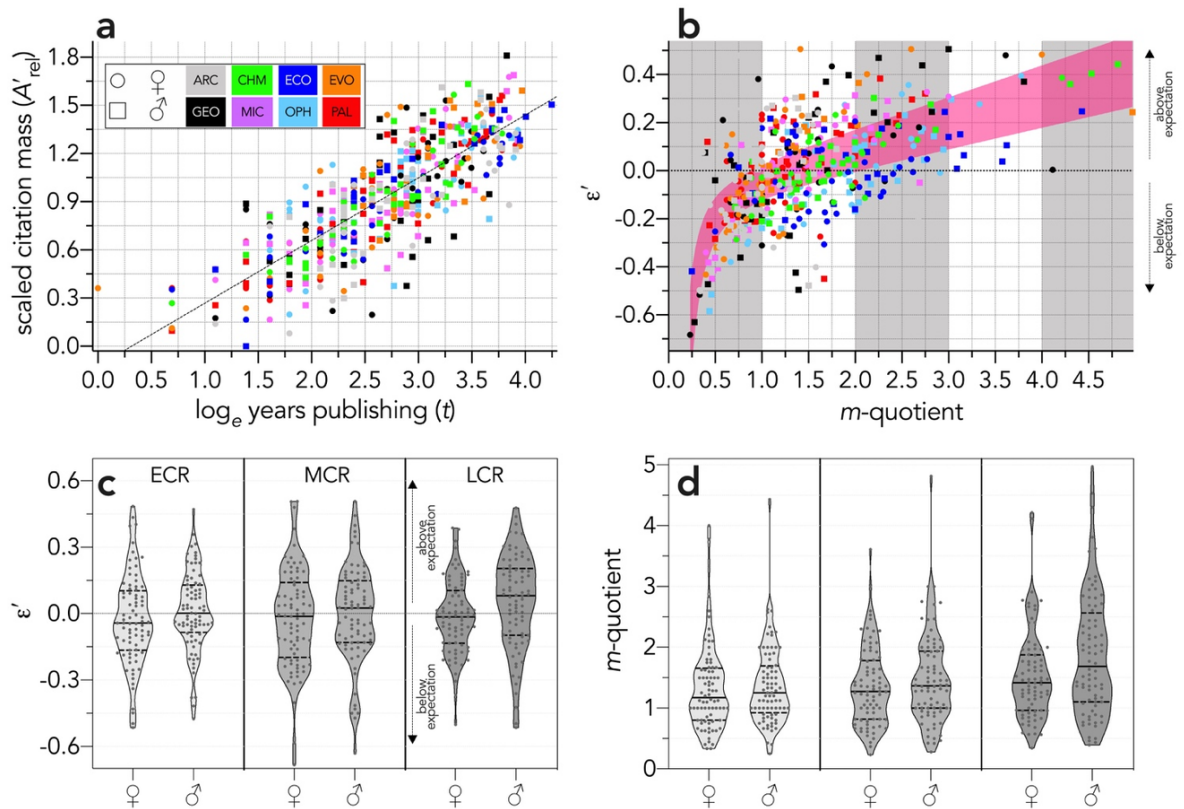


Figure 3: (a) Relationship between scaled citation mass (A'_{rel}) and \log_e years publishing (t) for 480 researchers in eight different disciplines (ARC = archaeology, CHM = chemistry, ECO = ecology, EVO = evolution and development, GEO = geology, MIC = microbiology, OPH = ophthalmology, PAL = palaeontology) comprising 60 researchers each (30 female, 30 male). (b) Relationship between the residual of $A'_{rel} \sim \log_e t$ (ϵ') and the m -quotient for the same researchers (pink shaded area is the 95% confidence envelope of a heat-capacity relationship of the form: $y = a + bx + c/x^2$, where $a = -0.17104 - 0.0875$; $b = 0.0880 - 0.1318$, and $c = -0.0423 - 0.0226$). (c) Truncated violin plots of ϵ' by gender and career stage (ECR = early career researcher, MCR = mid-career researcher, LCR = late-career researcher). When $\epsilon' < 0$, the researcher's citation rank is below expectation relative to her peers in the sample; when $\epsilon' > 0$, the citation rank is greater than expected relative to her peers in the sample (dashed lines = quartiles; solid lines = medians). (d) Truncated violin plot of the m -quotient by gender and career stage.

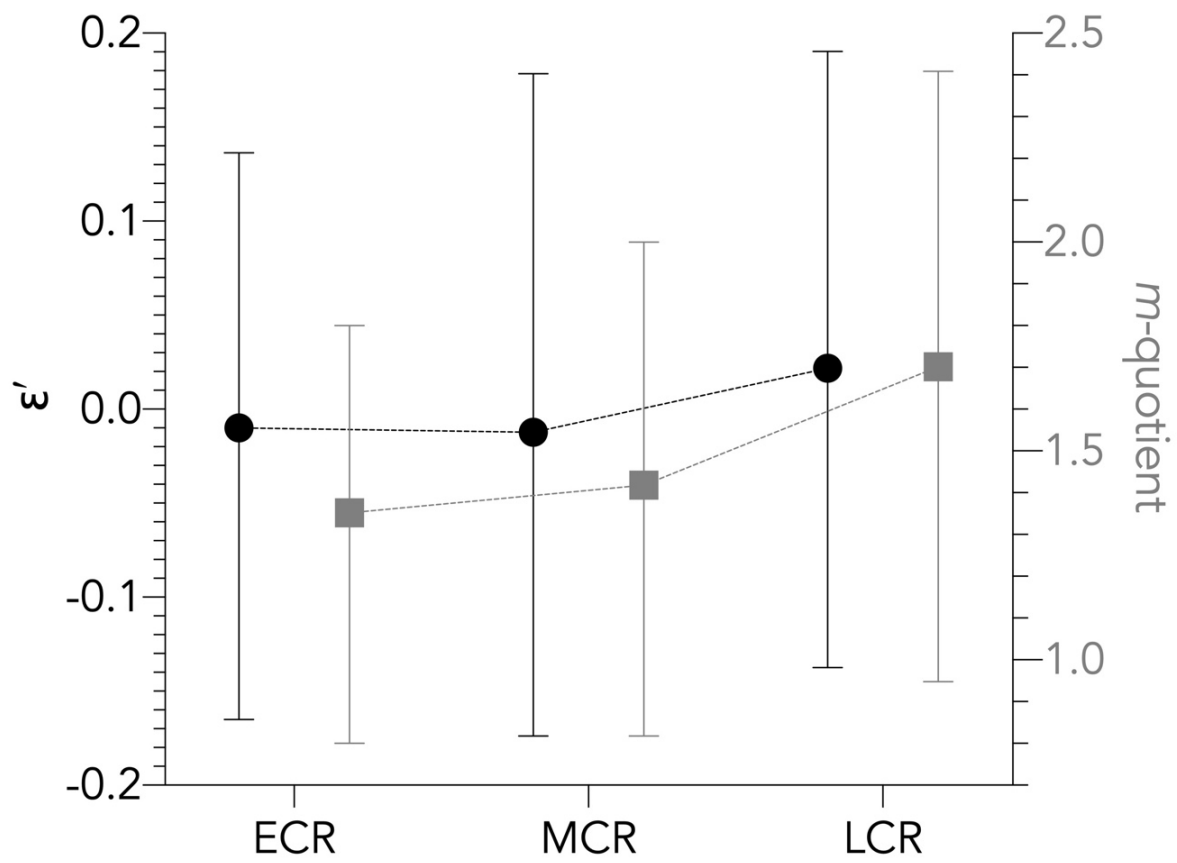


Figure 4: **Figure 4** | Median (with 80/20 percentiles) scaled residuals (ϵ' , black circles) and m -quotient (grey squares) across all eight disciplines relative to career stage (ECR = early career; MCR = mid-career; LCR = late career).

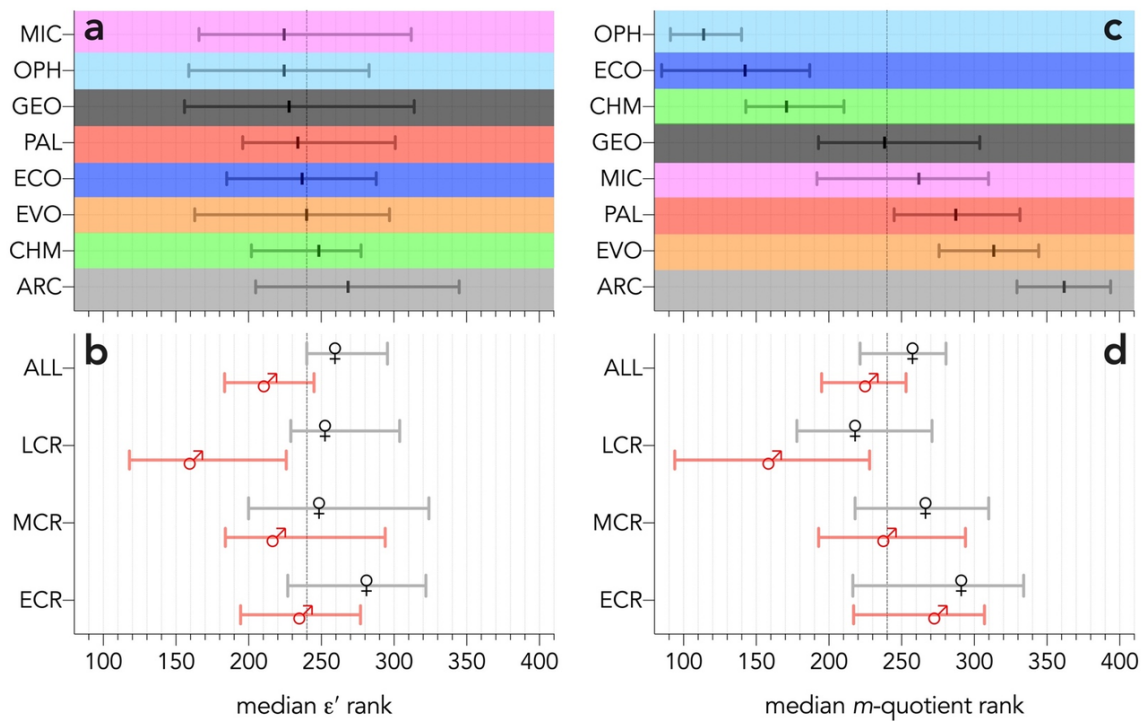


Figure 5: **Figure 5** | (a) Bootstrapped (10,000 iterations) median ranks among the eight disciplines examined (ARC = archaeology, CHM = chemistry, ECO = ecology, EVO = evolution and development, GEO = geology, MIC = microbiology, OPH = ophthalmology, PAL = palaeontology) based on the scaled residuals (ϵ'). (b) Bootstrapped ϵ' ranks by gender and career stage (ECR = early career researcher, MCR = mid-career researcher, LCR = late-career researcher). (c) Bootstrapped (10,000 iterations) median ranks among the eight disciplines based on the m -quotient. (d) Bootstrapped m -quotient ranks by gender and career stage. The vertical dashed line in all panels indicates the midway point across the entire sample (480 \div 2 = 240).

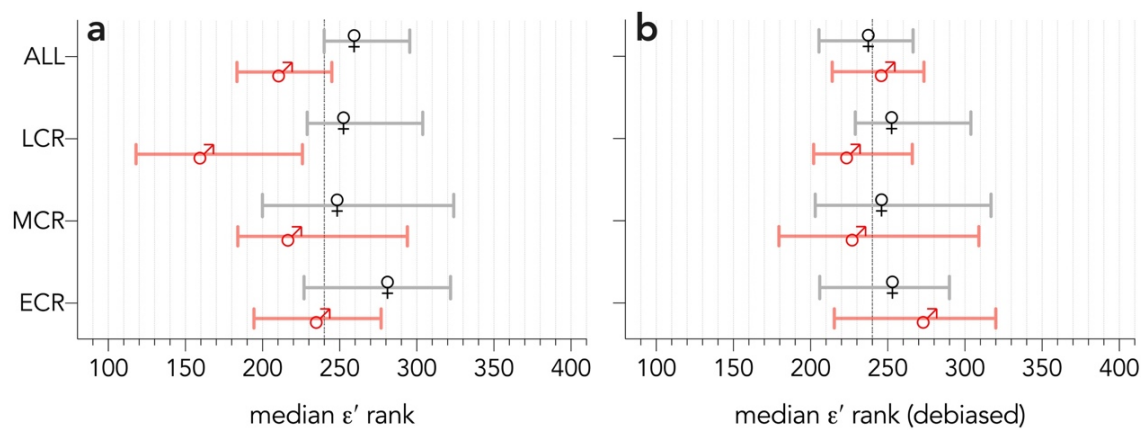


Figure 6: **Figure 6** | (a) Bootstrapped (10,000 iterations) ϵ' ranks by gender and career stage (ECR = early career researcher, MCR = mid-career researcher, LCR = late-career researcher); (b) bootstrapped debiased (i.e., calculating the scaled residuals for each gender separately, and then ranking the combined dataset) ϵ' ranks by gender and career stage.

Supplementary Information

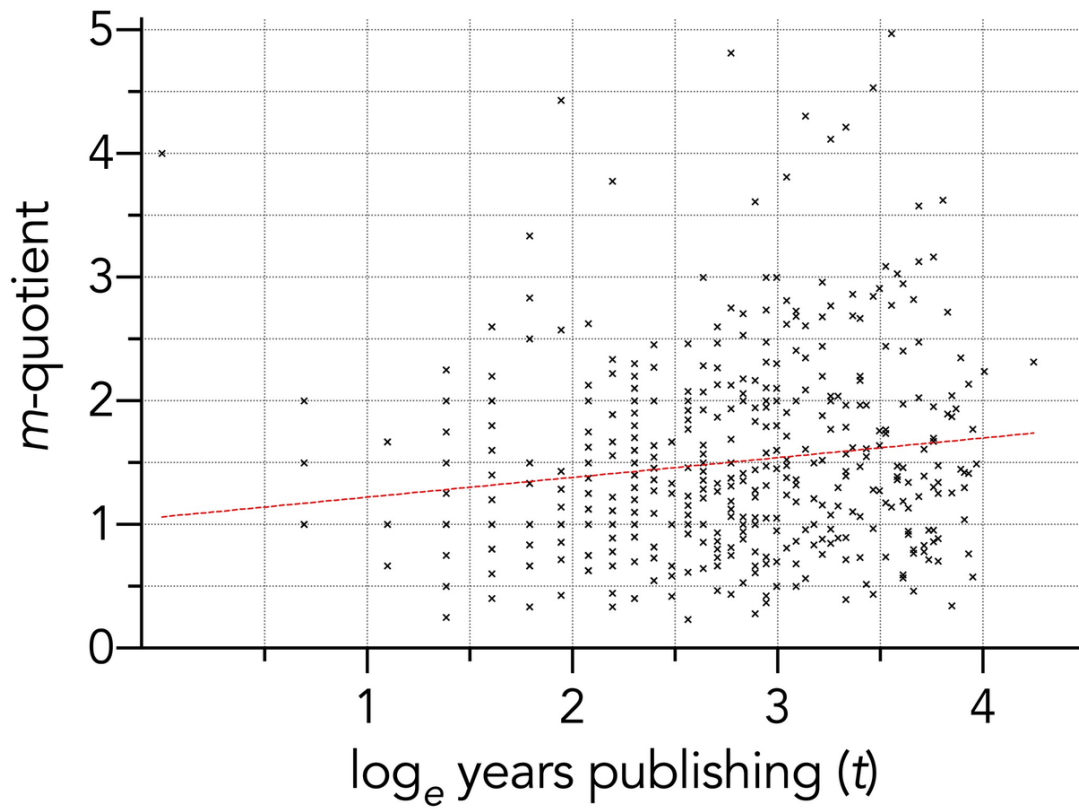


Figure 7: **Figure S1** | Relationship between the m -quotient and \log_e years publishing (t) for 480 researchers in eight different disciplines. There is a weak, but statistically supported positive relationship (information-theoretic evidence ratio = 68.7).

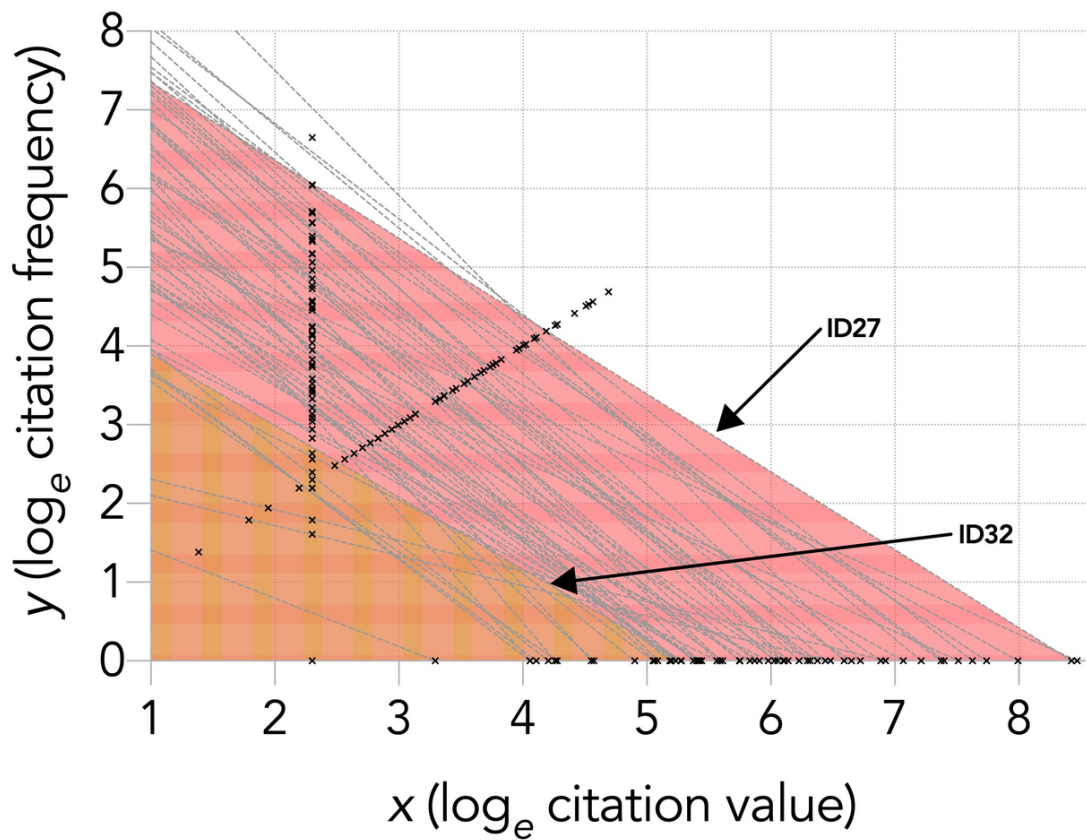


Figure 8: **Figure S2** | Relationship between log_e citation frequency (y) and log_e citation value (x) for 60 researchers within the discipline of ophthalmology. Each light grey, dashed line is the linear (on the log_e-log_e scale) fit for each individual researcher. The area under the fitted line (A_{rel}) is shown for individual 32 (ID32; red horizontal hatch) and individual 27 (orange vertical hatch).

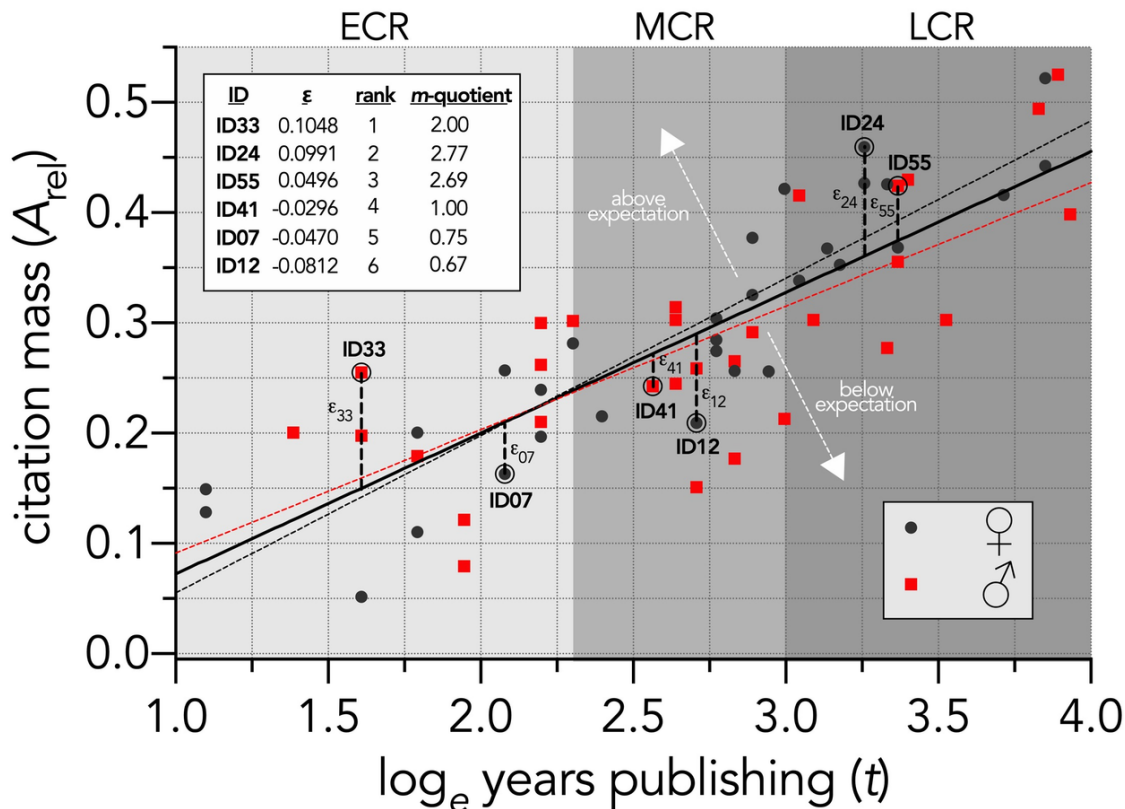


Figure 9: **Figure S3** | Relationship between a researcher's citation mass (A_{rel} ; area under the citation frequency–value curve — see Fig. S2) and \log_e years since first peer-reviewed publication (Y_1) for an example sample of 60 microbiology researchers in three different career stages: early career researcher (ECR), mid-career researcher (MCR), and late-career researcher (LCR). The residuals (ϵ) for each researcher relative to the line of best fit (solid black line) indicate relative citation rank — researchers below this line perform below expectation (relative to the sample), those above, above expectation. Also shown are the lines of best fit for women (black dashed line) and men (red dashed line — see also Fig. S4). Here we have also selected two researchers at random (1 female, 1 male) from each career stage and shown their results in the inset table. The residuals (ϵ) provide a relative rank from most positive to most negative. Also shown is each of these six researchers' m -quotient (h -index \div number of years publishing).

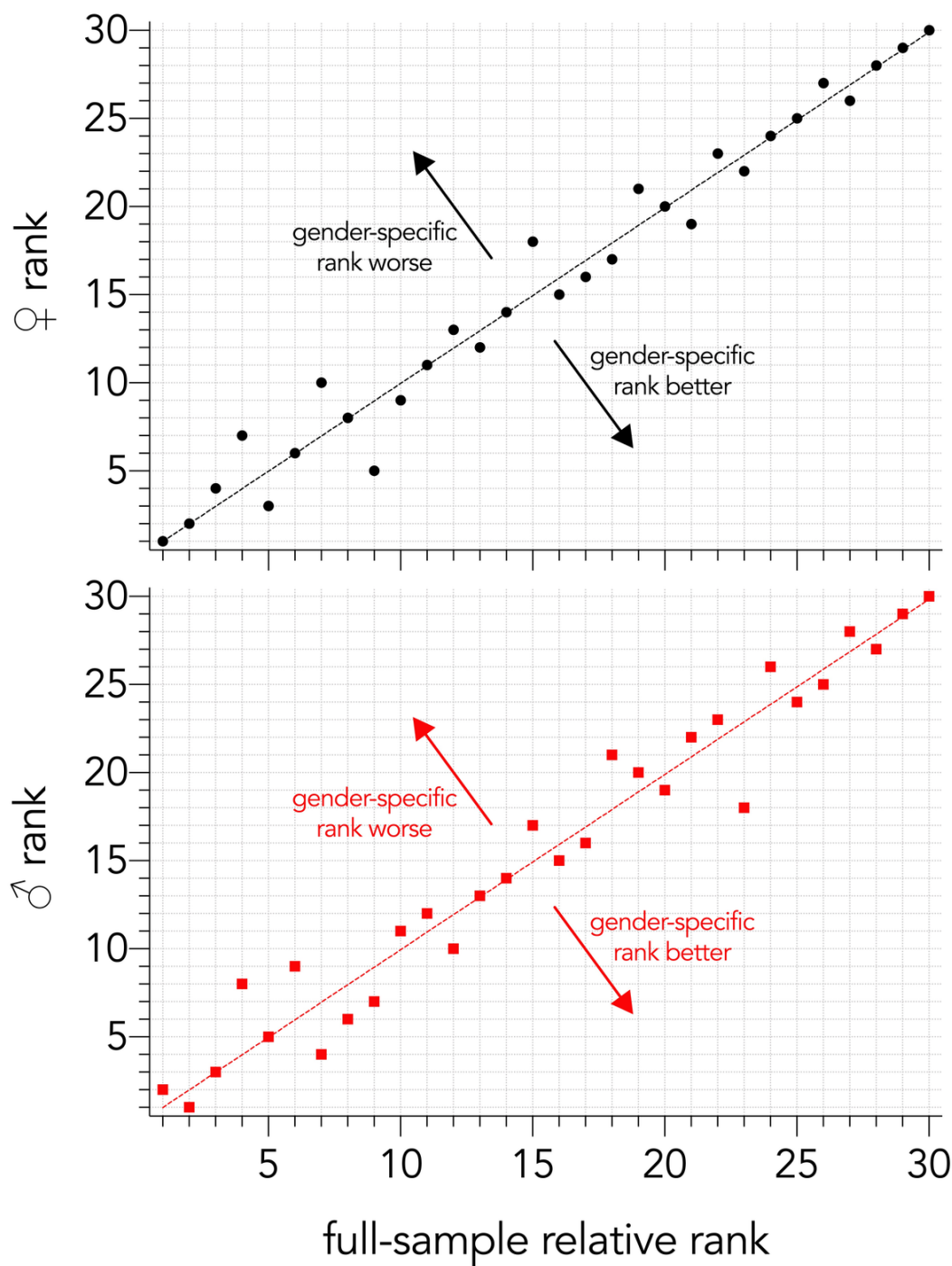


Figure 10: **Figure S4** | Gender-specific researcher ranks *versus* ranks derived from the entire sample (in this case, the microbiology sample shown in Fig. S3). For women who increased ranks when only compared to other women (negative residuals; top panel), the average increase was 1.50 places higher. For women with reduced ranks (positive residuals; top panel), the average was 1.88 places lower. For men who increased ranks when only compared to other men (negative residuals; bottom panel), or who declined in rank (positive residuals; bottom panel), the average number of places moved were both 1.75 for both.