



COMMENTARY

Is it possible to measure good science?

 Andrew N. Holding¹ , Kirsty R. McIntyre^{2,*}  and Paul T. Lynch^{3,†}

1 Department of Biology, University of York, York, YO10 5DD, UK

2 College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK

3 University Research and Knowledge Exchange Office, University of Derby, Derby, DE22 1GB, UK

Keywords

COVID-19; early career; equity; funding; impact; metrics; science

Correspondence

 A. N. Holding, Heads of University Biosciences (HUBS) Early Career Lecturers in Biosciences Advisory Group, Royal Society of Biology, 1 Naorji Street, London WC1X 0GB, UK
 Tel: 01904 328500
 E-mail: andrew.holding@york.ac.uk

Present address

 *Heads of University Biosciences (HUBS) Early Career Lecturers in Biosciences Advisory Group, Royal Society of Biology, 1 Naorji Street, London, WC1X 0GB, UK
 †Heads of University Biosciences (HUBS), Royal Society of Biology, 1 Naorji Street, London, WC1X 0GB, UK

(Received 26 August 2022, revised 24 October 2022, accepted 3 November 2022)

doi:10.1111/febs.16674

Metrics play a vital part in the valuation and funding of research for scientists worldwide. We review the challenges that metrics pose in providing a fair and equitable system for research funding. We highlight the attempts with declarations, including the San Francisco Declaration on Research Assessment (SF-DORA), to improve the research environment and specific impacts that metric choice can have on the evaluation and progression of Early Career Lecturers (ECLs). While there is much evidence that metrics will never be entirely satisfactory, we conclude there are opportunities that would benefit ECLs and reason for optimism for researchers.

Metrics are often a victim of their own success

Driving change is challenging. Not because people will not do as you ask, but often because people do exactly what you say. Policies that aim to monitor or alter the behaviour of others rarely cause people to alter their behaviour to achieve the aims of those who wrote the policy; instead, people may develop a habit of focusing on what they are being monitored on.

This observation is not new, you may have heard them in the form of Goodhart's Law. While the law was originally proposed to describe economics and monetary policy, it quickly became apparent that it applied far more widely. Today, you have probably heard it paraphrased as "When a measure becomes a target, it ceases to be a good measure" and even

Abbreviations

ECL, Early Career Lecturer; FWCI, Field-Weighted Citation Impact; ISE, Initiative of Science in Europe; JIF, Journal impact factor; KEF, Knowledge Exchange Framework; REF, Research Excellence Framework; SF-DORA, San Francisco Declaration on Research Assessment; TEF, Teaching Excellence Framework.

though Goodhart never wrote those words, it describes the challenges Goodhart chose to highlight well (Figure 1).

Metrics aren't all bad, or all good

Journal impact factor (JIF), the yearly average number of citations of articles published in the last 2 years in a given journal, is widely recognised to be a poor way of measuring the impact of researchers' work. Why should a paper that has been published in a journal that receives more citations mean that the work is necessarily higher impact? The answer is it does not, with impact factors in many top journals driven by a smaller number of highly cited papers offsetting the rest [1]. The citation counts for individual papers that influence JIF are complex too. The timing of publication in line with trending topics means some papers collect citations rapidly, as has been seen with recent COVID-19 publications, independent of the impact of the work. Similarly, papers that are either incorrect or have shortcomings may also have a high citation count, not because of their value, but because of subsequent reports highlighting its shortcomings.

The potential value of journal impact factors

The independent review of the work by the very people who are best placed to understand both the content and value is the benefit that underpins the peer-review process. For these very reasons, peer review is seen as the gold standard of assessing research output, and journals form a key part of that process. Given the work, time and effort that goes into the peer-review process of high-impact journals, one would hope that would

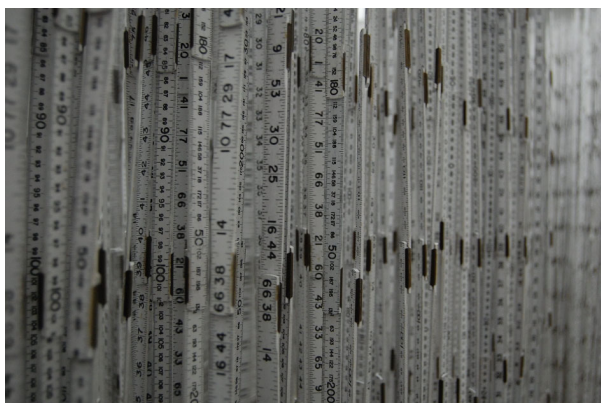


Fig. 1. Tape measures – Antony Mayfield (Licence CC: Attribution required) – <https://flic.kr/p/5UDjAw>.

translate into JIF as a metric of value. One cannot just ignore that piece of work has gone through a peer-review process before publication because of the challenge metrics face. These metrics have the potential to be especially valuable to support committee or decision-makers who do not have the specialisation or training to appraise the research. Yet, the link between good review and JIF is tenuous. Review is only part of the process leading to the publication and one cannot guarantee that all reviews are of high and similar quality. Further, JIF is very dependent on field, with different fields publishing papers of different scale and at different frequency. Researchers in fields like medicine and biochemistry typically produce more outputs with longer authors lists and can dominate non-weighted metrics; this is a reflection of research culture and not a measurement of productivity. Metrics like Field-Weighted Citation Impact (FWCI) do exist to manage this challenge; reflecting the performance of a paper within the research field rather than the journal in it was published in. A paper with a FWCI of 1 is cited as average for the field, while a FWCI of 1.5 is receiving 50% more citations. In this way, FWCI provides a potentially more accurate and nuanced way of comparing subject areas, although it is not perfect. FWCI still cannot directly measure research quality, and it is difficult to apply evenly to multi- and cross-disciplinary research.

The failure of journal impact factors in practice

High-quality peer review does not automatically translate into high-quality metrics either. An extreme example of the failure of JIF as a metric was that *Acta Crystallographica A* journal's impact factor jumped from under 3 to nearly 50 in 2009, all because of a single paper receiving over 5000 citations [2]. A sudden change in impact factor does not mean all the other papers in the same journal are more important. At the same time, appraising the quality of the research using these metrics provides an independent review of the work that, in some situations, is less subjective than internal reviews. The answer is not as simple as just scrapping metrics.

So important now is the value of a high-impact factor, it has on occasions distorted the aims of the journals too. In 2017, this was acutely felt when *Oncotarget* was delisted by several indexing services after apparent author coercion to include references back to their own journal. Overall, it was researchers who lost out, suddenly scientists found their work under unwanted scrutiny due to no fault of their own, with those at an earlier career stage with fewer papers being hit the hardest [3].

Journal impact factors are not the only problem

The *h*-index, defined as the maximum value *h* for which the author has *h* papers cited *h* times, is a perennial talking point. The idea is that your *h*-index measures the quality (number of citations) and quantity (number of papers) of your research output. As it is, the metric aims at providing an estimate of a scientist's standing within a single field, yet even Hirsch who proposed the index believes it can “fail spectacularly and have severe unintended negative consequences.” [4] Examples of this include driving academics to “hot topics” rather than those that benefit society to pick up citations more easily or encouraging scientists to game the system through self-citation. These concerns are based on genuine outcomes; a controversial promotion policy in Italy in 2017 led to a measurable increase in self-citation within the country [5]. While *h*-index is now generally accepted as a flawed measure of productivity, exemplified by the fact one can continue to increase their *h*-index well after their own death, the impact of metrics is still very much an important issue.

Early career vulnerability

The problem has not gone unrecognised, hiring committees and funding bodies are now much less likely to request your *h*-index, citation counts, or for you to put the impact factor next to journal names on your CV [6]. Yet the pervasiveness of metrics into research culture means that peer-review panels, grant boards and promotion panels are still influenced by them. Removing the impact factor of a journal does not mean that a “*Nature*” publication goes unrecognised on an early career CV. Nor does it make it easier to spot an important, but more obscure, piece of research that will go on to have a huge impact. The research behind CRISPR gene editing and mRNA vaccines both took years before the impact would be realised.

In the context of Early Career Lecturers (ECLs), these impacts are significant. Academic biology has a documented bias in those who make it to Professor at elite institutions [7], and citation counts often creep into the process of establishing the quality of research. Yet self-citation rate correlates strongly with gender, giving men an advantage [8].

Early Career Lecturers are additionally vulnerable to any solution put in place. If a metric is swapped or changed suddenly an ECL may find themselves no longer on the trajectory they once were. Should an ECL focus on a small number of high-impact papers,

or several smaller ones to increase the number of publications they have? What will help them progress their career? At the same time, these decisions have reduced impact for established researchers who will have a longer track record to fall back on. Ideally, all ECLs would have enough diversity not to notice such change, but with a sector that constantly reports on overwork and long hours, it is essential for ECLs to be strategic with their focus [9].

Post-COVID-19

With the impact of COVID-19 challenging the careers of ECLs, and established academic researchers, and clear data showing that men responded with a spike in publications not mirrored by women [10], there are concerns about how any metric will reflect this going forward. One thing is clear, the COVID-19 pandemic has amplified the inequities of the current system.

We must re-evaluate evaluation, we know funding cuts are likely, and we know that under the current research evaluation those cuts will hit minorities and women at the early career stage the hardest [11]. Countless examples during the pandemic have shown promising post-docs unable to balance full-time home-schooling with trying to take the steps in their career; particularly impacting women who are disproportionately likely to have responsibilities for child- or elder-care [12]. Those in an ECL post, but yet to secure their first big grant, may struggle in the next 12–36 months if successful funding forms a requirement as part of their probationary period. Whatever the career stage, the metrics used to evaluate individuals need to account for the challenges individuals have faced.

Better metrics

Early Career Lecturers are not alone in their concerns over the limitations of metrics like the *h*-index or JIF or the challenges they raise. Newer measures do exist, for example altmetric (<http://altmetric.com>) aims to capture the impact research may have beyond citation count, from policy documents to YouTube videos, enabling researchers to show the value of their research beyond citation count. As previously mentioned, FWCI was developed to enable better comparison across fields rather than the journal it was published but still, like JIF, it has limitations, including the potential of being influenced by highly-cited outlier publications. As time progresses, alternative career paths are forming within academia that recognise different types of academic output. For example,

the development of Research Software Engineers in response to the need to retain those with a combination of research and software skills who develop sustainable software in place of the traditional research publication. Research Software Engineers would not be sustainable if evaluated on the metrics of publications alone, yet meet an essential need in the research community.

Metrics also need to reflect the diversity of assessments; departmental- and institutional-level assessments have different aims and, therefore, different criteria. There is no “one size fits all” solution. For most individuals, the interactions in their early career are either via department recruitment panels or through the journey of probation and promotion. At recruitment, there is an opportunity to better recognise the role of individuals in delivering several multi-author studies versus a singular high-impact factor first authorship paper. This recognition is especially pertinent given that collaborations underpin a successful scientific career. There are also opportunities to include alternatives, for example, 360 reviews or reverse reviews, that can provide qualitative feedback to monitor skills such as mentorship that often go under-recognised [13].

Pushing for responsible metrics

While the impact of COVID-19 is very recent, the concerns on metrics are not. The 2013 San Francisco Declaration on Research Assessment (SF-DORA) is now signed by 145 countries, and over 2200 institutions. This declaration aims to find the balance between metrics to provide an accountable and transparent system, while highlighting the damage poorly designed ones, including JIF, can have. Fundamental to this strategy is that metrics should support the expert review not define it, yet there are still high-profile calls against that strategy [14].

The Initiative of Science in Europe’s (ISE) 2020 report on the precarity of academic careers “recommend[ed] assessors in the EU follow recent moves to reform research evaluation away from publication metrics” [15]. Yet despite SF-DORA and other attempts to improve metrics, including the Leiden Manifesto [16] and the Hong Kong principles [17], the report also noted “current approaches of assessment contradict this ideology.” Nonetheless, Research England’s current guidance overtly supports these aims with a specific expectation that “providers [they] fund will comply with the principles of SF-DORA, Leiden Manifesto or equivalent” in the current terms and conditions for UKRI grants [18].

REF and our early careers

The REF (Research Excellence Framework) plays a key part in providing accountability for research institutions like universities in the UK and informs the allocation of approximately £2 billion per year in public funding for university research. Within the guidance documents for REF, the use of metrics is specifically defined, and it states that metrics may only form part of the assessment in specific fields [19–21]. One of those fields is the biosciences, therefore metrics form a key part on how our research is rated, and the outcomes of the process are used to inform funding decisions going forward.

For ECLs, or those looking to become ECLs in the near future, understanding the implications of REF or comparable governmental research audits on hiring cycles is critical. Most notable is that, currently, hiring an ECL shortly before the completion of a REF cycle will enable the institution to use that ECL’s track record as part of their REF submission. However, the portability of research outputs is to be reviewed post-REF 2021 due to concerns that it leads to “poaching” of senior staff with good publications records just to boost REF metrics. Currently, there is no firm decision, but a complete move to non-portability would have a huge impact on employability of ECLs and is something to monitor over the coming years. The key decisions document for REF 2021 noted that “[non-portability] is an aspect of the policy that we will undoubtedly revisit when considering the arrangements for the next assessment exercise,” highlighting that many solutions resulted in significant unintended consequences on early career researchers [22].

For those of us currently in position, our own publication records have formed a key part in the Universities’ 2021 REF process with metrics playing a key part. The challenge of metrics is especially true for the biological sciences where citation data formed an indicator of the assessment for output quality and factual information on the significance of the output was not requested [23]. Concerns have not been ignored though, and the panel reviewing the citation data received guidance from the Forum for Responsible Research Metrics.

SF-DORA, the Leiden Manifesto and the Metric Tide [24] all have laid the groundwork and continue to push for responsible use of metrics (Figure 2). In particular, to stop the use of JIF as a proxy for the quality of individual research outputs, ensuring that we compare like-with-like and the use of a “basket” of metrics. The expected outcome is that using a collection of different measures of research impact, that is,

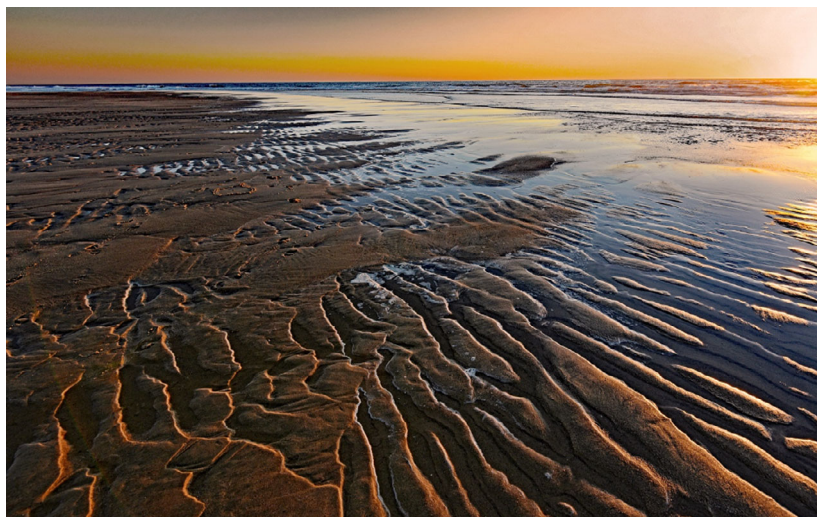


Fig. 2. Low Tide – Hervé Simon (Licence CC: Attribution required) – <https://flic.kr/p/2n9z6xJ>.

our basket, will prevent the hyperfocus and distortion that occurs with single metrics and provide opportunities for recognising more diverse outputs.

These aims present a challenge, those involved in REF submissions are often time-limited, not specialists in all fields they are to review, while the optimistic interpretation is that JIF reflects the outcome of multiple rounds of peer review. The solution should not be to fall back on JIF just because it is convenient, but we must acknowledge the alternative should not be overly time-consuming to measure and appraise within the context of academic workloads. Likewise, comparing like-with-like is a laudable aim and normalisation for the career stage and research field is clearly important. However, to what resolution does one capture this information?

The inclusion of a “basket” of metrics has merit and aims to prevent the problematic gaming of metrics by not applying too much pressure on a single outcome. The implementation of this strategy will also enable us to recognise the diversity and potential outcomes that are possible within the academic setting and reward institutions that support these endeavours. Some of these features are captured since March 2011 in the UKs annual Knowledge Exchange Framework (KEF) assessments, sitting alongside the REF and Teaching Excellence Framework (TEF), to measure the impact of research, learning/teaching and engagement in non-academic partnerships. The challenge here is to ensure that the basket is well-balanced and that the chosen metrics achieve what they have been selected for.

Promisingly, Research England has commissioned the “Metrics Tide Revisited” to review the current and potential uses of REF post-2021. The aim is not to

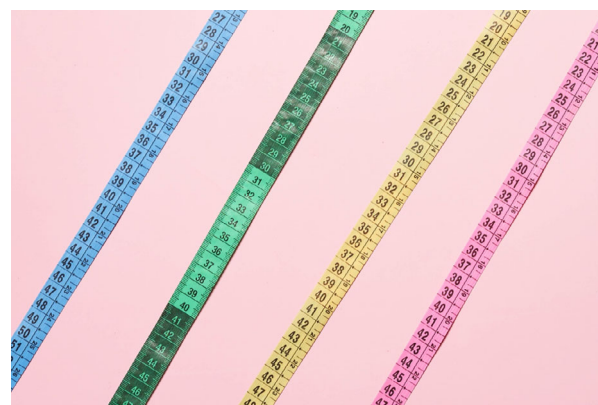


Fig. 3. Tape measures – Marco Verch (Licence CC: Attribution required) – <https://flic.kr/p/2kV3F8W>.

repeat the original study, as much as the original consolation work is still valid, but instead to monitor progress again at the conclusion of the initial report from 2015. While we cannot predict the outcomes of the review, those undertaking it are “quietly optimistic about the prospects for positive change” [25] (Figure 3).

Conclusion

Goodhart’s Law does not say that metrics are worthless, in fact their very ability to alter behaviour emphasises their value. It is the combination of measurement and metric combined that make them insidious, and after over a decade of complaining about *h*-indexes and JIFs, it is clear that they stubbornly refuse to go away. The solution is, instead of singularly focusing on the failures of metrics, to focus on continuing to

challenge the systems that drive the hyper-focus on them that corrupts their worth.

We should embrace proposals for a diversity of metrics and look to include alternatives that capture meaningful and often overlooked attributes that would improve the research environment beyond research output alone. The “toxic working culture” in universities has been discussed extensively in journals, blogs, and the media [26–29]. One potential response to the points that these articles raise is to change the assessment criteria to address pertinent issues, for example, the persistent undervaluing of the contributions from women and other minorities to research culture due to the prioritisation of metrics like JIF. A recent study of factors affecting women in STEM concluded that “women have “survived” their work environments despite structural barriers, only due to their determination, resilience, and fervent interest” [30]. Change is therefore much needed. Whatever those changes are; however, they should be designed with those they aim to support to ensure they meet the needs of those they aim to help.

There will always be pushback to these changes, solutions will undoubtedly mean those who did well under the previous schemes will do less well under the newer one. Nonetheless, these are not all-or-nothing solutions, we do not have to completely tear down the current system to be open to experimenting with alternative ways to summarise and understand the research output at institutional levels.

For ECLs, SF-DORA makes it explicitly clear that we should “not use journal-based metrics, such as Journal Impact Factors (JIFs), as surrogate measures of the quality of individual research articles, to assess an individual scientist’s contributions,” demonstrating the potential and value of these declarations. By implementing these changes, we have an opportunity for a more equitable system, moving away from short-term competition and towards quality research that may take longer or even fail. Something we should encourage because scientists are people, not just metrics on a page.

Acknowledgements

We thank the Royal Society of Biology Heads of University Biosciences (HUBS) for their role in creating and supporting the Early Career Lecturers in Bioscience (ECLBio) advisory group (Twitter: @ECL_HUBS).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

ANH conceived and prepared the initial draft of the manuscript. ANH, KRM and PTL commented on and wrote the final version of the manuscript. HUBS Early Career Lecturers in Biosciences Advisory Group edited and approved the final manuscript.

References

- 1 Editorial. Beware the impact factor. *Nat Mater*. 2013;**12**:89–91.
- 2 Dimitrov JD, Kaveri SV, Bayry J. Metrics: journal’s impact factor skewed by a single paper. *Nature*. 2010;**466**(7303):179–9.
- 3 [cited 2018 Mar 6]. Available from: <https://retractionwatch.com/2018/03/06/when-a-journal-is-delisted-authors-pay-a-price/>
- 4 [cited 2020 Mar 24]. Available from: <https://www.nature.com/nature-index/news-blog/whats-wrong-with-the-h-index-according-to-its-inventor>
- 5 [cited 2018 Jun 4]. Available from: <https://www.natureindex.com/news-blog/italian-scientists-increase-self-citations-in-response-to-promotion-policy>
- 6 [cited 2022 Apr 31]. Available from: <https://www.ukri.org/about-us/research-england/research-excellence/research-metrics/>
- 7 [cited 2014 Jun 30]. Available from: <https://news.mit.edu/2014/research-reveals-gender-gap-nations-biology-labs-0630>
- 8 Chawla DS. Men cite themselves more than women do. *Nature*. 2016;**535**(7611):212.
- 9 Powell K. Young, talented and fed-up: scientists tell their stories. *Nature*. 2016;**538**:446–9.
- 10 [cited 2020 May 14]. Available from: <https://github.com/drfrfeder/pandemic-pub-bias/blob/master/README.md>
- 11 [cited 2021 Mar 24]. Available from: <https://www.timeshighereducation.com/news/sir-paul-nurse-ukri-cuts-are-existential-threat-science>
- 12 Editorial. COVID is amplifying the inadequacy of research-evaluation processes. *Nature*. 2021;**12**:89–91.
- 13 McCarthy AM, Garavan TN. 360 Feedback process: performance, improvement and employee career development. *J Eur Ind Train*. 2001;**25**:5–32.
- 14 [cited 2021 Mar 20]. Available from: <https://www.timeshighereducation.com/news/creator-says-ref-should-swap-expert-panels-metrics-science>
- 15 ISE Task Force on Researchers’ Careers. Position on precarity of academic careers. Strasbourg: Initiative for Science in Europe; 2020.
- 16 Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols I. Bibliometrics: the Leiden Manifesto for research metrics. *Nature*. 2015;**520**(7548):429–31.
- 17 Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong principles for

- assessing researchers: fostering research integrity. *PLoS Biol.* 2020;**18**(7):e3000737.
- 18 [cited 2021 Aug 1]. Available from: <https://re.ukri.org/sector-guidance/publications/terms-and-conditions-of-research-england-grant-2021-22/>
- 19 [cited 2022 Jul 14]. Available from: <https://www.timeshighereducation.com/news/ref-related-funding-rise-10-cent-ps2-billion-year>
- 20 [cited 2020 Jul 29]. Available from: <https://www.ref.ac.uk/publications/guidance-on-revisions-to-ref-2021/>
- 21 [cited 2020 Jul 29]. Available from: <https://www.ref.ac.uk/guidance/citation-and-contextual-data-guidance/>
- 22 [cited 2019 Jan 30]. Available from: <https://www.ref.ac.uk/media/1080/key-decisions-by-the-ref-steering-group.pdf>
- 23 [cited 2019 Mar 1]. Available from: <https://www.ref.ac.uk/publications/panel-criteria-and-working-methods-201902/>
- 24 Wilsdon J. The metric tide: independent review of the role of metrics in research assessment and management. Thousand Oaks, CA: SAGE Publications Ltd; 2016.
- 25 [cited 2022 May 17]. Available from: <https://www.researchprofessionalnews.com/rr-news-political-science-blog-2022-5-the-metric-tide-rises-again/>
- 26 [cited 2019 Oct 10]. Available from: <https://www.universityaffairs.ca/opinion/the-black-hole/research-culture-in-biomedical-science-needs-to-change/>
- 27 Editorial. A kinder research culture is possible. *Nature.* 2019;**574**:5–6.
- 28 Nik-Zainal S. Research culture must be kinder for all, not just royalty. *Nature.* 2019;**575**:287.
- 29 Muscatelli A. Universities must overhaul the toxic working culture for academic researchers. *Guardian.* 2020; [cited 2020 Jan 15]. Available from: <https://www.theguardian.com/education/2020/jan/15/universities-must-overhaul-the-toxic-working-culture-for-academic-researchers>
- 30 Prieto-Rodriguez E, Sincock K, Berretta R, Todd J, Johnson S, Blackmore K, et al. A study of factors affecting womens lived experiences in STEM. *Hum Soc Sci Commun.* 2022;**9**(1):1–11.