

Sample size in bibliometric analysis

Gordon Rogers¹ · Martin Szomszor¹ · Jonathan Adams^{1,2}

Received: 11 May 2020 / Published online: 31 July 2020 © The Author(s) 2020

Abstract

While bibliometric analysis is normally able to rely on complete publication sets this is not universally the case. For example, Australia (in ERA) and the UK (in the RAE/REF) use institutional research assessment that may rely on small or fractional parts of researcher output. Using the Category Normalised Citation Impact (CNCI) for the publications of ten universities with similar output (21,000-28,000 articles and reviews) indexed in the Web of Science for 2014-2018, we explore the extent to which a 'sample' of institutional data can accurately represent the averages and/or the correct relative status of the population CNCIs. Starting with full institutional data, we find a high variance in average CNCI across 10,000 institutional samples of fewer than 200 papers, which we suggest may be an analytical minimum although smaller samples may be acceptable for qualitative review. When considering the 'top' CNCI paper in researcher sets represented by DAIS-ID clusters, we find that samples of 1000 papers provide a good guide to relative (but not absolute) institutional citation performance, which is driven by the abundance of high performing individuals. However, such samples may be perturbed by scarce 'highly cited' papers in smaller or less research-intensive units. We draw attention to the significance of this for assessment processes and the further evidence that university rankings are innately unstable and generally unreliable.

Keywords Bibliometric sampling \cdot CNCI \cdot Citation impact \cdot Research assessment \cdot University ranking

Introduction

What is the minimum number of observations required to make an acceptably precise estimate of the true mean citation impact or describe the relative means of a number of datasets? Sampling to estimate the population mean is a widespread problem in many research areas (e.g. Adams, 1980), but it is less commonly an issue when estimating citation impact in bibliometrics because it is often possible to make use of complete

The Policy Institute, King's College London, 22 Kingsway, London WC2B 6LE, UK



[☐] Jonathan Adams jonathan.adams@kcl.ac.uk

Institute for Scientific Information, Clarivate Analytics, 160 Blackfriars Road, London SE1 8EZ, IJK

data, i.e. the full publication set for one or more entities. Of course, we make the caveat that this is a complete dataset only insofar as it is complete for a particular source such as the *Web of Science*. Other but unrecorded publications usually exist.

Circumstances may arise in research assessment where the analysis of all available publication data will not or cannot be the case, in which event some light needs to be shed on sampling acceptability. We have sought to explore this because it is a challenge posed to us by many users of bibliometric analysis. It is a truism that larger samples reduce the variance of the mean but at what sample size does differentiation between a series of datasets become sufficiently clear to satisfy evaluation requirements?

Moed et al. (1985) analysed the consequences of operating on incomplete bibliometric data in research evaluation. They concluded that "a completeness percentage of 99% for publication data is proposed as a standard in evaluations of the performance of small university research groups". This seems a high boundary but Glaser et al. (2004) also refer to the 'least evaluable unit (LEU)' in organisational research assessment and comment that "the main obstacles to further disaggregation below the LEU are that indicators lose their statistical validity because of low numbers of publications and that the performance of subunits cannot be independently measured". Calatrava Moreno et al. (2016) found that "indicators of interdisciplinarity are not capable of reflecting the inaccuracies introduced by incorrect and incomplete records because correct and complete bibliographic data can rarely be obtained".

Glänzel and Moed (2013) have discussed the issue of indicator consistency. They note that "as a rule of thumb a value of 50 is suggested as minimum value for approximate properties such as 'normality' of the distribution of means and relative frequencies. In [a worked example of Belgian bibliometric data], a sample size of the order of magnitude of 100 was used and provided acceptable results." Seglen (1994) studied the consistency of the relationship between article citedness and journal impact for Norwegian biomedical researchers. He found that "very large numbers of articles (50–100) had to be pooled in order to obtain good correlations" so groups above the author level were obligatory. Shen et al. (2019) do present a method for estimating minimum sample size for accurate bibliometric ranking but their algorithm is applicable to paired data-sets, exemplified with journals having similar impact factors.

It is generally true that much larger samples are available to analysts but exceptions occur, particularly where quantitative and qualitative research evaluation are linked in practical assessment processes. This may then influence the number of outputs available for analysis. For example, in the Australian Excellence in Research for Australia (ERA), consideration of a sample size that might be judged adequately representative of a unit's work led the 2008 Indicators Development Group to recommend a low volume threshold (ERA 2018) that was set at 50 assessable outputs. A different situation is found where there is intentional selectivity. The numbers of publications that are reviewed in the UK's Research Assessment Exercise (RAE, later the Research Excellence Framework or REF) is limited by practical considerations of the reasonable workload for a peer review panel. Hitherto, each RAE assessable researcher has submitted four outputs from their portfolio over a census period of several years (HEFCE 2014) but this system is changing. For the next REF there must be a minimum of one output for each submitted researcher plus further outputs up to a multiple of 2.5 for the submitted staff count with a maximum of five outputs attributed to any individual (REF 2019). This 'pick-and-mix' could have significant consequences for different units according to the staff balance.

The notion of sampling and representation requires some comment in this context. Research evaluation may emphasise both the proportion of activity that is 'excellent' and



the 'average' performance of a unit (Glänzel and Moed (2013) refer to this as 'the high end and the common run'). For example: in the RAE, research managers are likely to want to submit outputs that represent excellence and it is assumed that they select outputs that represent researchers' most impactful research (Adams et al. 2020). Similar assumptions are made in Brazilian research assessment (Capparelli and Giacomolli 2017). In ERA, by contrast, the intention is to capture the typical performance of the unit.

A third context in which partial samples may cover only some of the assessed unit's activity is that of university rankings. This is an area of some sensitivity regarding precision and sample size, either in a national exercise or in a wider, global context. If the variance of possible outcomes is high then the likelihood that the relative performance and status of institutions would be misinterpreted may make the reporting unacceptable. While this seems unlikely at the level of major research institutions, it could be the case for subject-based analyses and may affect specialist institutions where a relatively large fraction of output is not in indexed journals. It may also affect the analysis of less research-intensive institutions that collaborate in larger, global studies that produce a small number of highly cited papers.

It is obvious from basic statistical theory that larger samples lead to analytical outcomes that are likely to be more 'accurate' in the sense of providing a result that is closer to the true population mean. However, given the nature of citation distributions, which are invariably highly skewed (Seglen 1992), is it possible to determine a reasonable 'practical' threshold for a minimum acceptable sample size? What happens when the structure of sampling is determined by the researchers themselves? There has been little work on this since the general intention of evaluation has been, as noted above, to capture as much information as possible rather than to limit analysis to samples. However, a guide to the general relationship between sample size and outcomes may be of value in guiding policy and implementation for national and institutional exercises, and to avoiding erroneous assumptions about representativeness.

Because clients have frequently raised the question of sample size with the Institute for Scientific Information (ISITM) we have considered the question of how an indicator of an institution's citation performance might be affected by partial analysis of its output. To do this, we analysed the Category Normalised Citation Impact (CNCI) of a set of comparable institutions and asked two questions regarding the use of partial data:

- At what sample size would the variance from true average CNCI invalidate interpretation of relative outcomes (the ERA scenario)?
- If we intentionally sample more highly cited items for researchers, how does this affect the variance and relative status of outcomes (the REF scenario)?

Methods

Data are drawn from the *Web of Science* Core Collection using the Science Citation Index Expanded, the Social Science Citation Index and the Arts and Humanities Citation Index (SCIE, SSCI and AHCI) for the 5-year period 2014–2018. Documents for analysis were restricted to original academic journal contributions (i.e. articles and reviews) which we will refer to as 'papers'.

Because citation counts grow over time at rates that are field-dependent (Garfield 1979), we calculate Category Normalised Citation Impact (CNCI) for each individual paper. This



Region	Institution	Number of papers	Average CNCI	Rank CNCI	% cited
Europe	ETH Zurich	27,869	1.78	2	92.7
Asia	Nanjing University	27,567	1.37	7	90.1
Europe	University of Edinburgh	25,965	2.04	1	88.4
Americas	University of Texas, Austin	25,205	1.69	4	88.3
Asia	Yonsei University	24,362	1.12	8	87.1
Americas	University of Montreal	23,701	1.51	5	88.2
Europe	University of Helsinki	23,420	1.76	3	89.3
Americas	UNAM	22,112	0.83	10	81.7
Russia	Lomonosov Moscow State University	22,083	0.94	9	78.6
Middle East	Tel Aviv University	21,788	1.43	6	85.2

Table 1 Paper counts (article and review), average Category Normalized Citation Impact (CNCI) and citedness of ten university institutions, ranked by volume of output using data from the *Web of Science Core Collection* for the 5-year period 2014–2018

takes into account the average citation count for all the papers in a subject-based category of journals and for their year of publication.

We also use the arithmetic mean as a standard indicator although we are aware of the well-founded recommendations of Thelwall (2016; see also Fairclough and Thelwall 2015) regarding the use of the geometric mean for these skewed data. However, noting Thelwall's comments regarding palatibility to policy-makers and given the practical context within which this is to be applied, we believe that arithmetic means are sufficiently satisfactory and intuitively more accessible for practical purposes.

To provide a comparative group of academic institutions, ten universities of a similar output size with a wide geographic spread were selected. The aim was to assure sampling comparability by identifying institutions that produced a similar output count of about 20k–30k papers over 5 years, of which there are 59. From this pool it was possible to select: three from the Americas; three from Europe excluding Russia; two from Asia–Pacific; and one each from Russia and the Middle East. The actual size boundaries of the institutions selected, which ranged from 21,000 to 28,000 papers during the 5-year period, should provide a sound basis for comparability.

The range of CNCI values in each of these institutional portfolios is, of course, both very great and very skewed, with many uncited papers and low citation values and a long tail of much higher CNCI values (Glänzel 2013). However, the question is not the precision of the institutional average but the degree to which our sampling scenarios can provide an informative estimate of that average (sample size) or of the ordinal relationship between institutional averages (highly cited papers).

Limiting sample size

To examine the variance due to sample size, we use simple random sampling (without replacement) to extract 10,000 different samples across a range of samples sizes. The completed sample is then replaced. This means that, for any sample, the total pool of papers to be drawn from is the same and each paper can only be selected once. These samples sizes were 20, 50, 100, 200, 500, 1,000 and 2000 papers, providing a range from approximately



0.1% to 10% of the total population. For each sample, the mean CNCI was calculated using baselines derived from *Web of Science* category, year of publication, and document-type.

Selecting highly cited papers

To examine the variance due to selective choices, we used the researcher-specific clusters created by Clarivate's Distinct Author Identification System (DAIS). DAIS uses a weighted comparison of author clusters drawing on over twenty points of distance/similarity from publication metadata including the author's ORCID, name, subject category, use of references, author-based co-citation analysis (Small 1973; White and Griffith 1981), institutional name and so on. Detail of related methodology is in Levin et al. (2012). The system as applied to *Web of Science Core Collection* data also responds to user feedback to improve aggregation and separation. It is regularly tested by ISI with manual verification drawing on disambiguated and validated Highly Cited Researcher records, which indicate 99.9% precision and 95.5% recall.

For each set of papers attributed to an institution, we extract all DAIS-IDs (i.e. clusters of papers that are associated with a unique researcher) and select the paper with the highest CNCI for each. This creates a set of 'top' papers produced by the institution. Many of the most highly cited papers have multiple co-authors from the same institution so these are de-duplicated (i.e. only included once). Only DAIS-IDs with 4 or more papers were included to filter out any potential outliers that were not properly disambiguated or have low publication output. This filtered sub-set results in around 3000 DAIS-IDs for each institution. As this is a 5-year time frame, most clusters had less than 10 papers and there were few that exceeded 25 papers at any institution. The distribution of these 'top' CNCI values is illustrated in Appendix (Fig. 8) and discussed below.

Simple random sampling was used to extract 10,000 samples of 1000 papers for each institution.

Results

Limiting sample size

Because of the large number of sampling iterations, the average of the mean CNCIs from the samples was similar to the population CNCI for all institutions at all sample sizes. At all sample sizes, the distribution of means typically approached a normal distribution. The skew, which even at small sample sizes was much less than the source population skew, rapidly decreased. Some institutions had a bimodal distribution, which is discussed separately below.

The statistic of interest is not the average value of a large sample and its departure from the population mean but the variance in the sample means (Fig. 1).

The variance associated with small sample sizes is very high (Fig. 1). The range of variances is correlated with the average CNCI of the institution, which is obviously a derivative of the distribution of individual paper CNCIs. No institution has a uniformly highly cited set of papers but the spread (kurtosis) of its individual paper CNCI values is greater where the institution's average CNCI is higher. Helsinki and Edinburgh have two of the three highest average CNCIs and have a relatively platykurtic (though skewed) distribution with a wide range of individual paper CNCIs from which samples



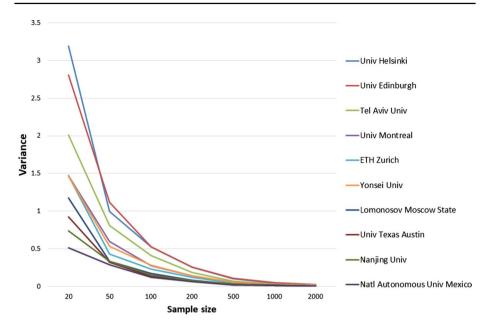


Fig. 1 Variance in the mean value of CNCI calculated from 10,000 iterative samples of papers (articles and reviews) taken from the complete *Web of Science Core Collection* 5 year (2014–2018) publication set for ten universities of similar portfolio size (see Table 1)

may be drawn. UNAM has a low average CNCI and a clustered range of paper CNCIs (more leptokurtic) because it has relatively smaller number of highly cited papers.

The variance was greater than 1 for seven and over 0.5 for all the institutions at sample size = 20 papers. It dropped to a range up to around 1.0 at sample size = 50, and to 0.5 or less at sample size = 100. It can be seen that an increase in the chosen range of sample sizes broadly halves the variance at each step (Fig. 1).

At what points on this spectrum do the ranges of sample CNCI values broadly overlap and at what point does the variance drop to a level such that the probability that the sample value is approaching the true CNCI suggests that the universities can be more accurately distinguished?

In Fig. 2, the population average CNCI values for the full dataset of papers for the ten institutions are shown with an indicator of the magnitude of the standard deviation (which should cover slightly more than two-thirds of the datapoints) at each of three sample sizes. It is evident that a sample size of 50 produces a relatively high probability of indistinguishable results. In this scenario, the ranking of institutions by CNCI could vary considerably

Even with sample sizes of 200 there is an appreciable likelihood of misinterpretation. If we consider Tel Aviv University, with an average CNCI near the middle of our institutional set, then we can see that a spread of other institutional means from Yonsei to Zurich lie within the range of one standard deviation. In fact, the ranges of all the institutional standard deviations still overlap except the institutions with the two lowest and three highest average CNCIs. The institutions in the middle range of mean CNCIs are effectively indistinguishable at this level of sampling.



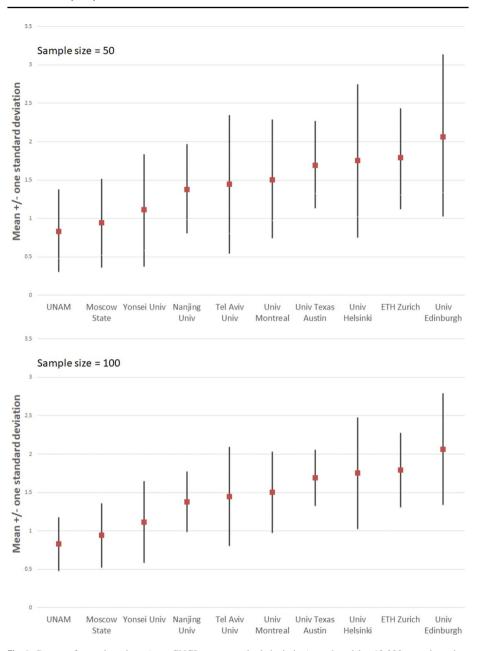


Fig. 2 Range of sample values (true $CNCI\pm$ one standard deviation) produced by 10,000 samples taken from the full publication set for ten universities for samples of 50, 100, 200 and 500 papers

Selecting highly cited papers

A naive expectation, when the highest impact papers (by CNCI) are selected from each DAIS-ID cluster, would be that the smaller dataset created by removing low-cited



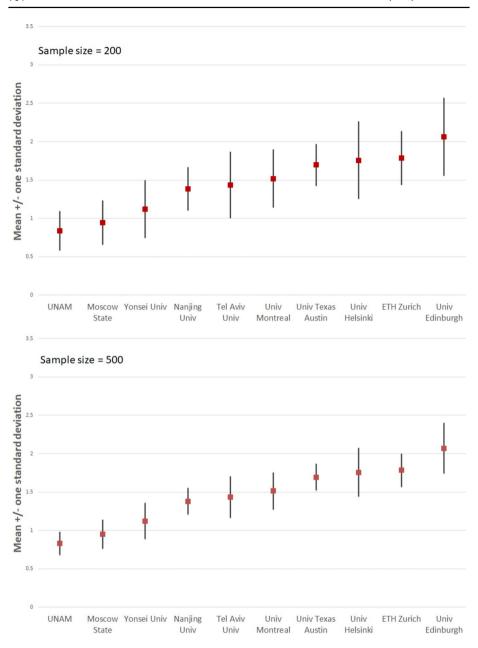


Fig. 2 (continued)

papers would lead to an increase in the average CNCI for an institution and the variance would similarly go down because low-cited papers have been removed. Reality does not match this at aggregated institutional level, however, because of the variance between DAIS-IDs, some of which are mostly highly cited and some of which are mostly low cited, particularly among social sciences and arts outside North America. We therefore



reduced the dataset and took the paper with the highest CNCI from each cluster, as described in Methods.

For information, the overall institutional distributions of the subset of researchers' 'top' CNCI papers for DAIS-IDs with four or more papers was plotted for each institution (Appendix: Fig. 7). The spread of most impactful papers (in terms of CNCI) for the set of researchers at these institutions are skewed in a similar way to the overall CNCI distribution. It is interesting, however, to note the similarity of distribution between many institutions, with modal CNCI values around 1–2 times world average and a tail extending into the 4–8 times world average. Indeed, institutional differences in this tail may be a principal differentiator (Glänzel 2013).

There is a general agreement in the scientometric literature that, on average, there is a broad relationship between average CNCI values and other quantitative (research income) and qualitative (peer review) indicators of research performance (reviewed in Waltman 2016). Figure 7 (in Appendix) therefore seems to suggest that the researcher population at each institution is made up of a very large platform of common-run individuals (sensu Glänzel and Moed 2013) whose most highly cited papers are a little above world average and a right-skewed tail of high-end researchers whose papers are much more highly cited for their field and year. The relative distribution of the mainstream and the talented must then influence the net institutional outcome.

Because the population is so skewed, the standard deviation and hence the error in the sample means also increase. The main driver for this is the residual skewed distribution: although some low cited papers have been removed, there are still plenty of other low cited papers. The removal of various low cited papers leads to an increase in the mean, the remaining low-citation papers are now further from the mean as a consequence, and therefore the standard deviation is greater. Although the average CNCI of the highest CNCI papers in each DAIS-ID cluster is correlated with and about 2.5 to 3 times the overall mean CNCI for each institution (Fig. 3), the means are statistically indistinguishable for the distributions of researchers' highest CNCI papers (Appendix: Fig. 8).

There were about 3000 papers (range Moscow—2392 to Nanjing—4195) in the 'top' papers' dataset for each institution. Sampling this dataset, using 10,000 iterations of 1000 papers each, produces the aforementioned increase in the average CNCI for each institution, since many low-cited papers have been removed. The distributions of sample means are shown in Fig. 4. Although the underlying distribution remains very skewed (Appendix: Fig. 7) the distribution of the sample means is much smaller and again approaches normality.

UNAM's distribution in Fig. 4 has a double peak and is clearly not normal. Lomonosov Moscow University may also have an emergent second peak. Further investigation (below) was carried out to explore the source of this anomaly.

The distributions of sample means can be seen to be relatively discrete (Fig. 5) and provides a better level of discrimination than did the samples of 200 papers from the full population (Fig. 2). As noted earlier, given that the modal peaks of highest CNCI values are similar across these institutions (Appendix: Fig. 7) the differentiating factor that separates the much tighter peak values must be the relative frequency of higher 'top' CNCI values (see Glänzel 2013). Thus, in research assessment exercises where selectivity is supported, the ability to select such material will be of critical significance in determining the outcomes.

It is feasible to analyse the data based on direct analysis of indicative author names, but no substantive difference in the results is provided by this.



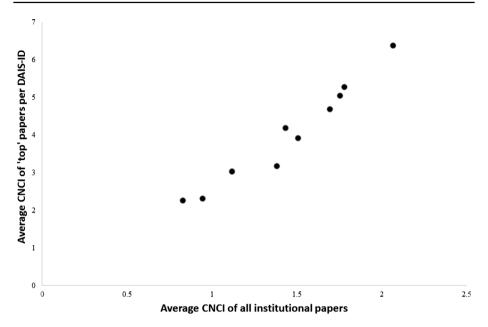


Fig. 3 Relationship between the average CNCI of all institutional papers indexed in the *Web of Science Core Collection* for a 5-year period (2014–2018) and the average CNCI of the most impactful ('top') paper in each DAIS-ID cluster with four or more papers

Bi-modal distributions

While the distribution of sample means (for the full institutional data and for the 'top' papers data) was typically normal, some of the institutions had a double peak in the distribution of their sample means, particularly for larger sample sizes. This was investigated by progressive sampling of the UNAM data (which has the most evident bimodality) with a greater number of sample size intervals from very small (20 papers) to comprehensive (2000 papers) samples from the 'top' paper dataset of 2714 papers for the 5-year period.

Figure 6 shows a plot of the distributions resulting from this spread of varying sample sizes using 10,000 samples at each interval. The horizontal axis shows the range of average CNCI for the samples and was set to a maximum of 10 times world average since a valid institutional average greater than this would be extremely unlikely. The distribution appears unimodal with a very small sample size of 20 because the right-hand modal peak is in fact above a CNCI of 20 and is thus out-of-frame. As the sample size is increased to 50 it just begins to come into view on the right of the plot. As the eye progresses through increasing sample sizes it is evident that this peak grows in frequency and moves leftward.

Why does this happen? It is a consequence of one UNAM paper being particularly highly cited compared with the rest of the institution's output. Samples that included this paper would, of course, have a distinctly higher mean CNCI, while the probability that this paper was included in a sample is a simple function of the sample size. Samples with and without this paper separately approached normal distributions but when combined produce a double peak. The most highly cited UNAM paper has a CNCI of ~418; its second most highly cited paper has a CNCI of ~91, followed by progressively closer CNCIs of 73, 70 and 54. The peak on the right in Fig. 6 represents the samples that include that highly cited paper whereas the peak on the left denotes the samples that where it was not captured. The



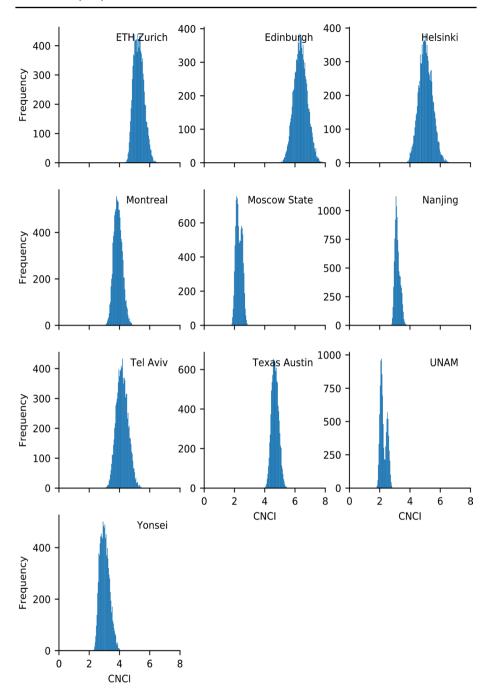


Fig. 4 Frequency distributions of 10,000 sample mean CNCI for 1000 papers taken from the highest CNCI paper in institutional DAIS-ID sets > 4 papers (see text)

peak on the right should be centred around $(418 - \bar{x})/n$, where \bar{x} is the mean of the left peak (around 2) and n is the sample size. For a sample size of 1000, the peak would be



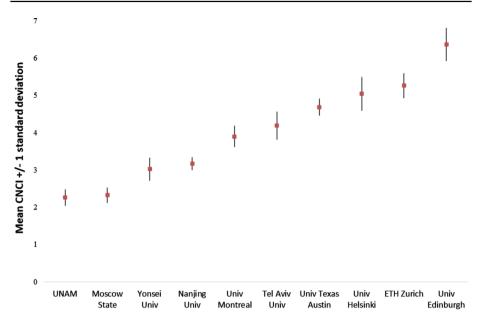


Fig. 5 Range of sample values (mean CNCI±one standard deviation) produced by 10,000 samples of 1000 papers sampled from the highest CNCI papers for DAIS-ID sets for ten universities

0.416 to the right of the other peak while the standard deviation of either peak is 0.22, and, with 2714 papers in the top-cited papers dataset, the right-hand peak should have a height of n/(2,714-n) compared with the left peak (i.e. the probability that a sample of n randomly chosen papers includes that top paper against the probability that it doesn't).

Discussion

The scenarios and questions posed here are purely experimental. In practice, it is unlikely that anyone would be so ill advised as to seek to compare the citation performance of a global set of institutions on what are obviously relatively small sample sizes or on a single highly cited paper per researcher. The experiment does, however, throw some light on how policymakers should consider reasonable limits to sampling bibliometric data and what advice analysts might offer users. It might also enhance the cautionary approach to interpretation of any analysis that compared relatively similar groups or institutions.

Our experimental notion of 'sampling' is based on a select group of institutions with similar and relatively large publication portfolios (20,000–30,000 papers over 5 years). If we had chosen smaller and larger institutions then the source population sizes would have been a further interactive factor. We know, from practical experience, that the citation indicators for smaller institutions (and even some small countries) can be surprisingly influenced by atypically highly cited papers (Potter et al. 2020).

It is no surprise that the variance around average institutional CNCI is very high when small sample sizes are employed and that this drops rapidly as sample size increases (Fig. 1). A sample size of 50, and even of 100, still produces an outcome with a weak likelihood of accurately identifying and differentiating the true CNCI values for our set of institutions. Only when the sample size reaches 200 does an appreciable degree of



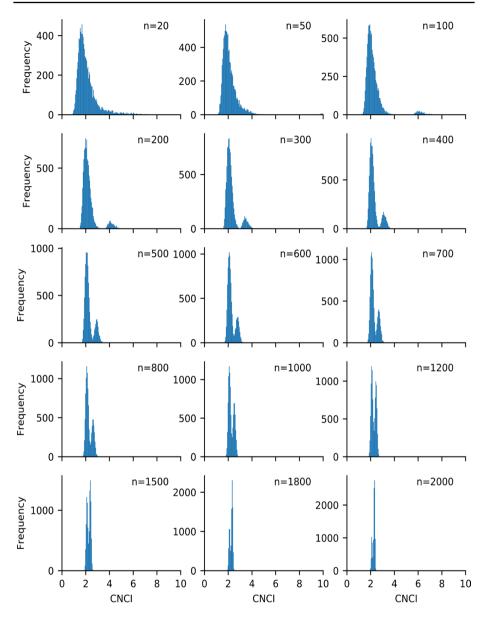


Fig. 6 The distributions of mean CNCI for 10,000 samples of 2714 'top' papers from DAIS-ID clusters for UNAM researchers, with sample sizes from 20 to 2000 papers. The horizontal axis (mean CNCI) is restricted to a maximum institutional value of 10 times world average. A 'second peak', caused by a very small number of very highly cited papers, can be seen in samples above 50 and progressively moves leftwards with increasing sample size



differentiation begin to appear (Fig. 2). From the perspective of a national assessment exercise, a sample as small as 50 (as used in ERA) would appear to be insufficient were any bibliometric analysis to be employed but that is not to say that it would not provide valuable information to an expert and experienced peer review panel.

The focus on the higher CNCI papers for each researcher (as represented by DAIS-ID clusters) revealed the extent to which the distribution of 'top' papers for the institutional populations is as skewed as that of the citation performance of the papers themselves (see also Glänzel 2013). There are high frequencies of relatively low cited researcher clusters, even in leading institutions with a high average CNCI, and so sampling across researchers produces distributions with a high average but an increased variance because the average has moved away from even the best papers of the low cited. This pattern is captured by Glänzel and Moed (2013) in their reference to 'the high end and the common run'.

In application, the sample distributions of researchers' 'top' CNCI papers were well defined (Fig. 4) and produced a differentiation between institutions that is as good as large sample sizes (Fig. 5). The key factor driving the average CNCI of these researcher 'top' papers is the relative abundance of the more highly cited researchers, since the modal values are similar for all ten institutions analysed here (Appendix: Fig. 7).

The influence on the sample average of scarce papers with exceptional citation counts is shown by the UNAM data analysed in Fig. 6. The likelihood that such papers are included in a sample depends on both relative abundance within the institutional portfolio and the size of the sample. A small sample would therefore risk double jeopardy for the analyst where an institution has few such papers. Most samples would miss such rare items, but the average CNCI of a sample that included such a paper would be extremely—even absurdly—high.

The shift in research assessment methodology in the UK to allow different numbers of submitted outputs per researcher (REF 2019) will likely produce analytical outcomes that will depend significantly on local strategies. Institutions that adopt an inclusive approach, where all submitted researchers are equally represented, will tend towards lower average CNCIs whereas those that adopt an exclusive approach favouring the research leaders and the more highly cited will tend to elevate their relative position, though possibly at some cost to collegiality. This would not have been evident under the historical process which required an equal number of outputs for every submitted researcher. It is fortunate that debates in the UK have led to the decision to make use of bibliometrics only in some panels and then only in a peripheral, background manner.

Finally, there is the question of international rankings and the comparative position of institutions in such rankings. It should now be more evident than before that even very large samples may not adequately differentiate between the many institutions in the centre of the performance distribution. The elite are likely to be well differentiated and a tail of research-sparse institutions may also be clear. In the middle, however, the degree of database coverage, the subject portfolio and other factors are likely to produce outcomes which would see an institution move up or down by many points each year. Even the completeness



and accuracy with which authors describe their affiliation may influence this. Such variation may be partly addressed by using 'bands' rather than ordinal points, but ultimately the only way of judging an institution's relative value is by detailed consideration of the underlying evidence.

Acknowledgements We thank our ISI colleagues for their advice and suggestions during the development of this work.

Compliance with ethical standards

Conflict of interest The authors are employees of the Institute for Scientific Information (ISI), which is a part of Clarivate, the owners of the *Web of Science* Group.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix

See Figs. 7, 8.



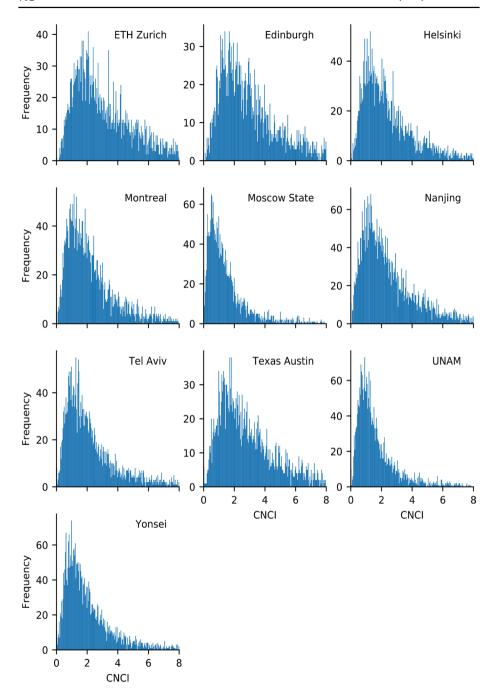


Fig. 7 Distribution of highest CNCI papers in each DAIS-ID with four or more papers



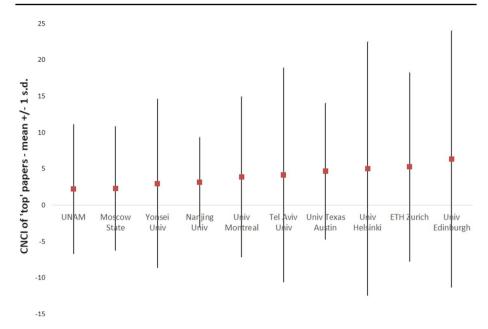


Fig. 8 Range of impact values (average CNCI±one standard deviation) for the distribution of highest CNCI ('top') papers taken from DAIS-ID clusters with four or more papers for each of ten universities

References

Adams, J. (1980). The role of competition in the population dynamics of a freshwater flatworm Bdellocephala punctata (Turbellaria, Tricladida). Journal of Animal Ecology, 49, 565–579.

Adams, J., Gurney, K. A., Loach, T., & Szomszor, M. (2020). Evolving document patterns in UK research assessment cycles. Frontiers in Research Metrics and Analytics, 5, 2.

Calatrava Moreno, M. D. C., Auzinger, T., & Werthner, H. (2016). On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. *Scientometrics*, 107(1), 213–232. https://doi.org/10.1007/s11192-016-1842-4.

Capparelli, B., & Giacomolli, N. J. (2017). The evaluation of impact factor in the scientific publication of criminal procedure. Revista Brasileira de Direito Processual Penal, 3(3), 789–806. https://doi. org/10.22197/rbdpp.v3i3.108.

ERA. (2018). Excellence in research for Australia: Submission Guidelines, p. 72, © Commonwealth of Australia 2017. ISBN: 978-0-9943687-4-4 (online).

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics*, 9(4), 895–906. https://doi.org/10.1016/j.joi.2015.09.005.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? Scientometrics, 1(4), 359–375. https://doi.org/10.1007/BF02019306.

Glänzel, W. (2013). High-end performance or outlier? Evaluating the tail of scientometric distributions. Scientometrics, 97(1), 13–23. https://doi.org/10.1007/s11192-013-1022-8.

Glänzel, W., & Moed, H. F. (2013). Opinion paper: Thoughts and facts on bibliometric indicators. Scientometrics, 96(1), 381–394. https://doi.org/10.1007/s11192-012-0898-z.

Glaser, J., Spurling, T. H., & Butler, L. (2004). Intraorganisational evaluation: Are there 'least evaluable units'. Research Evaluation, 13(1), 19–32.

HEFCE. (2014). REF2014: Assessment criteria and level definitions. http://www.ref.ac.uk/2014/panels/assessmentcriteriaandleveldefinitions/. Last accessed April 06, 2020.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, 63(5), 1030–1047.



- Moed, H., Burger, W., Frankfort, J., & Van Raan, A. (1985). The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics*, 8(3), 177–203. https://doi.org/10.1007/BF02016935.
- Potter, R. W. K., Szomszor, M., & Adams, J. (2020). Interpreting CNCIs on a country-scale: The effect of domestic and international collaboration type. *Journal of Informetrics*, 14(4), 101075.
- REF. (2019). Guidance on submissions. Research excellence framework 2019/01. https://www.ref.ac.uk/publications/guidance-on-submissions-201901/. Last accessed April 15, 2020.
- Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, 43(9), 628–638.
- Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, 45(1), 1–11. https://doi.org/10.1002/(sici)1097-4571(19940 1)45:1%3c1:aid-asi1%3e3.0.co;2-v.
- Shen, Z., Yang, L., Di, Z., & Wu, J. (2019). Large enough sample size to rank two groups of data reliably according to their means. *Scientometrics*, 118(2), 653–671. https://doi.org/10.1007/s11192-018-2995-0.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269.
- Thelwall, M. (2016). The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. *Journal of Informetrics*, 10(1), 110–123. https://doi.org/10.1016/j.joi.2015.12.001.
- Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007.
- White, H. D., & Griffith, B. C. (1981). Author co-citation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 32, 163–171.

